

PCE Working Group  
Internet Draft  
Intended status: Standard Track  
Expires: January 14, 2014

Zafar Ali  
Siva Sivabalan  
Clarence Filsfils  
Cisco Systems

Robert Varga  
Pantheon Technologies

Victor Lopez  
Oscar Gonzalez de Dios  
Telefonica I+D

July 15, 2013

Path Computation Element Communication Protocol (PCEP)  
Extensions for remote-initiated GMPLS LSP Setup  
draft-ali-pce-remote-initiated-gmpls-lsp-01.txt

Status of this Memo

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

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

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

This Internet-Draft will expire on January 13, 2014.

Copyright Notice

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

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

Expires January 2014

[Page 1]

## Abstract

PCEP Extensions for PCE-initiated LSP Setup in a Stateful PCE Model draft [I-D. draft-crabbe-pce-pce-initiated-lsp] specifies procedures that can be used for creation and deletion of PCE-initiated LSPs under the active stateful PCE model. However, this specification is focused on MPLS networks, and does not cover remote instantiation of GMPLS paths. This document complements PCEP Extensions for PCE-initiated LSP Setup in a Stateful PCE Model draft by addressing the extensions required for GMPLS applications, for example for OTN and WSON networks.

When active stateful PCE is used for managing PCE-initiated LSP, PCC may not be aware of the intended usage of the LSP (e.g., in a multi-layer network). PCEP Extensions for PCE-initiated LSP Setup in a Stateful PCE Model draft does not address this requirement. This draft also addresses the requirement to specify on how PCC should use the PCEP initiated LSPs.

## Conventions used in this document

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC 2119 [RFC2119].

## Table of Contents

1. Introduction.....	3
2. Use Cases.....	4
2.1. Single-layer provisioning from Active stateful PCE.....	4
2.2. Bandwidth-on-demand for multi-layer networks.....	5
2.3. Higher-layer signaling trigger.....	6
2.4. NMS-VNTM cooperation model (separated flavor).....	8

Expires January 2014

[Page 2]

3. GMPLS Requirements for Remote-Initiated LSPs.....	9
4. Remote Initiated LSP Usage Requirement.....	10
5. PCEP Extensions for Remote-Initiated GMPLS LSPs.....	10
5.1. Generalized Endpoint in LSP Create Message.....	11
5.2. GENERALIZED-BANDWIDTH object in LSP Create Message.....	11
5.3. Protection Attributes in LSP Create Message.....	12
5.4. ERO in LSP Create Object.....	12
5.4.1. ERO with explicit label control.....	12
5.4.2. ERO with Path Keys.....	13
5.4.3. Switch Layer Object .....	13
6. PCEP extension for PCEP Initiated LSP Usage Specification....	14
6.1. LSP_TUNNEL_INTERFACE_ID Object in LSP Create Message....	14
6.2. Communicating LSP usage to Egress node.....	15
6.3. LSP delegation and cleanup .....	16
7. Security Considerations.....	16
8. IANA Considerations.....	16
8.1. END-POINT Object.....	16
8.2. PCEP-Error Object.....	16
9. Acknowledgments.....	16
10. References.....	16
10.1. Normative References.....	16
10.2. Informative References.....	17

## 1. Introduction

The Path Computation Element communication Protocol (PCEP) provides mechanisms for Path Computation Elements (PCEs) to perform route computations in response to Path Computation Clients (PCCs) requests. PCEP Extensions for PCE-initiated LSP Setup in a Stateful PCE Model draft [I-D. draft-ietf-pce-stateful-pce] describes a set of extensions to PCEP to enable active control of MPLS-TE and GMPLS tunnels.

[I-D. draft-crabbe-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 that is centrally controlled and deployed. However, this specification is focused on MPLS networks, and does not cover the GMPLS networks (e.g., WSON, OTN, SONET/ SDH, etc. technologies). GMPLS requirements for PCEP initiated LSPs are outlined in Section 3. This document complements [I-D. draft-crabbe-pce-pce-initiated-lsp] by addressing the requirements for remote-initiated GMPLS LSPs. The PCEP extensions for PCEP initiated GMPLS LSPs are specified in Section 5. The mechanism described in this document is applicable not only to active PCEs initiating LSPs, but to any entity that initiates LSPs remotely.

When an active stateful PCE is used for managing remote-initiated LSP, the PCC may not be aware of the intended usage of the remote-initiated LSP. For example, the PCC may not know the target IGP instance in which the remote-initiated LSP is to be used. These requirements are outlined in Section 4. [RFC6107] defines LSP\_TUNNEL\_INTERFACE\_ID Object for communicating target IGP instance and usage of the forwarding and/ or routing adjacency from the ingress node to the egress node. However, current PCEP specifications do not include signaling of the LSP\_TUNNEL\_INTERFACE\_ID TLV in the PCEP message. Furthermore, [I-D. draft-crabbe-pce-pce-initiated-lsp] does not address this requirement. This draft also addresses the requirement to specify on how PCC should use the PCEP initiated LSPs. This is achieved by using LSP\_TUNNEL\_INTERFACE\_ID Object defined in [RFC6107] in PCEP, as detailed in Section 6.

## 2. Use Cases

### 2.1. Single-layer provisioning from active stateful PCE

Figure 1 shows a single-layer topology with optical nodes with a GMPLS control plane. In this scenario, the active PCE can dynamically create or delete L0 services between client interfaces. This process can be triggered by the deployment of a new network configuration or a re-optimization process. This operation can be human-driven (e.g. through an NMS) or an automatic process.

[Please refer to pdf version for the Figure]

Figure 1. Single-layer provisioning from active stateful PCE.

L0 PCE obtains resources information via control plane collecting LSAs messages. The request contains, at least, two optical transport interfaces (OT i/f), so PCE computes the path and sends a message to the optical equipment with ERO path information.

## 2.2. Bandwidth-on-demand for multi-layer networks

This use case assumes there is a multi-layer network composed by routers and optical equipment. In this scenario, there is an entity, which decides it needs extra bandwidth between two routers. This certain moment a GMPLS LSP connecting both routers via the optical network can be established on-the-fly. This entity can be a router, an active stateful PCE or even the NMS (with or without human intervention).

It is important to note that the bandwidth-on-demand interfaces and spare bandwidth in the optical network could be shared to cover many under capacity scenarios in the L3 network. For example, in this use-case, if we assume all interfaces are 10G and there is 10G of spare bandwidth available in the optical network, the spare bandwidth in the optical network can be used to connect any router, depending on bandwidth demand of the router network. For example, if there are three routers, it is not known a priori if the demand will make bandwidth-on-demand interface at R1 to be connected to bandwidth-on-demand interface at R2 or R3. For this reason, bandwidth-on-demand interfaces cannot be pre-provisioned with the IP services that are expected to carry.

According to [RFC5623], there are four options of Inter-Layer Path Computation and Inter-Layer Path Control Models: (1) PCE-VNTM cooperation, (2) Higher-layer signaling trigger, (3) NMS-VNTM cooperation model (integrated flavor) and (4) NMS-VNTM cooperation model (separated flavor). In all scenarios there is a certain moment when entities are using an interface to request for a path provisioning. In this document we have selected two use cases in a scenario with routers and optical equipment to obtain the requirements for this draft, but it is applicable to the four options.

[Please refer to pdf version for the Figure]

Figure 2. Use case higher-layer signaling trigger

Internet-Draft     draft-ali-pce-remote-initiated-gmpls-lsp-01.txt

### 2.3. Higher-layer signaling trigger

Figure 2 depicts a multi-layer network scenario similar to the presented in section 4.2.2. [RFC5623], with the difference that PCE is an active stateful PCE [I-D. draft-ietf-pce-stateful-pce].

In this example, O1, O2 and O3 are optical nodes that are connected with router nodes R1, R2 and R3, respectively. The network is designed such that the interface between R1-O1, R2-O2 and R3-O3 are setup to provide bandwidth-on-demand via the optical network.

The example assumes that an active stateful PCE is used for setting and tearing down bandwidth-on-demand connectivity. Although the simple use-case assumes a single PCE server (PCE1), the proposed technique is generalized to cover multiple co-operating PCE case. Similarly, although the use case assumes PCE1 only has knowledge of the L3 topology, the proposed technique is generalized to cover multi-layer PCE case.

The PCE server (PCE1) is assumed to be receiving L3 topology data. It is also assumed that PCE learns L0 (optical) addresses associated with bandwidth-on-demand interfaces R1-O1, R2-O2 and R3-O3. These addresses are referred by OTE-IP-R1 (optical TE link R1-O1 address at R1), OTE-IP-R2 (optical TE link R2-O2 address at R2) and OTE-IP-R3 (optical TE link R3-O3 address at R3), respectively. How PCE learns the optical addresses associated with the bandwidth-on-demand interfaces is beyond the scope of this document.

How knowledge of the bandwidth-on-demand interfaces is utilized by the PCE is exemplified in the following. Suppose an application requests 8 Gbps from R1 to R2 (recall all interfaces in Figure 1 are assumed to be 10G). PCE1 satisfies this by establishing a tunnel using R1-R4-R2 path. PCEP initiated LSP using techniques specified in [I-D. draft-crabbe-pce-pce-initiated-lsp] can be used to establish a PSC tunnel using the R1-R4-R2 path. Now assume another application requests 7 Gbps service between R1 and R2. This request cannot be satisfied without establishing a GMPLS tunnel via optical network using bandwidth-on-demand interfaces. In this case, PCE1 initiates a GMPLS tunnel using R1-O1-O2-R2 path (this is referred as GMPLS tunnel1 in the following). The PCEP initiated LSP using techniques specified in document are used for this purpose.

As mentioned earlier, the GMPLS tunnel created on-the-fly to satisfy bandwidth demand of L3 applications cannot be pre-provisioned in IP network, as bandwidth-on-demand interfaces and spare bandwidth in the optical network are shared. Furthermore, in this example, as active stateful PCE is used for managing PCE-initiated LSP, PCC may not be aware of the intended usage of the PCE-initiated LSP. Specifically, when the PCE1 initiated GMPLS tunnel1, PCC does not know the IGP instance whose demand leads to establishment of the GMPLS tunnel1 and hence does not know the IGP instance in which the GMPLS tunnel1 needs to be advertised. Similarly, the PCC does not know IP address that should be assigned to the GMPLS tunnel1. In the above example, this IP address is labeled as TUN-IP-R1 (tunnel IP address at R1). The PCC also does not know if the tunnel needs to be advertised as forwarding and/ or routing adjacency and/or to be locally used by the target IGP instance. Similarly, egress node for GMPLS signaling (R2 node in this example) may not know the intended usage of the tunnel (tunnel1 in this example). For example, the R2 node does not know IP address that should be

assigned to the GMPLS tunnel1. In the above example, this IP address is labeled as TUN-IP-R2 (tunnel IP address at R2). Section 6 of this draft addresses the requirement to specify on how PCC and egress node for signaling should use the PCEP initiated LSPs.

#### 2.4. NMS-VNTM cooperation model (separated flavor)

Figure 3 depicts NMS-VNTM cooperation model. This is the separated flavor, because NMS and VNTM are not in the same location.

A new L3 path is requested from NMS, because there is an automated process in the NMS or after human intervention. NMS does not have information about all network information, so it consults L3 PCE. For shake of simplicity L3-PCE is used, but any other multi-layer cooperating PCE model is applicable. In case that there are enough resources in the L3 layer, L3-PCE returns a L3 only path. On the other hand, if there is a lack of resources at the L3 layer, the response does not have any path or may contain a multilayer path with L3 and L0 (optical) information in case of a ML-PCE. In case of there is not a path in L3; NMS sends a message to the VNTM to create a GMPLS LSP in the lower layer. When the VNTM receives this message, based on the local policies, accepts the suggestion and sends a similar message to the router, which can create the lower layer LSP via UNI signaling in the routers, like in use case in section 2.3.1. Similarly, VNTM may talk with L0-PCE to set-up the path in the optical domain (section 2.2). This second option looks more complex, because it requires VNTM configuring inter-layer TE-links.

Requirements for the message from VNTM to the router are the same than in the previous use case (section 2.3.1). Regarding NMS to VNTM message, the requirements here depends on who has all the information. Three different addresses are required in this use case: (1) L3, (2) L0 and (3) inter-layer addressing. In case there is a non-cooperating L3-PCE, information about inter-layer connections have to be stored (or discovered) by VNTM. If there is a ML-PCE and this information is obtained from the network, the message would be the same than in section 2.3.1.

[Please refer to pdf version for the Figure]

Figure 3. Use case NMS-VNTM cooperation model



Expires January 2014

[Page 8]

### 3. GMPLS Requirements for Remote-Initiated LSPs

[I-D. draft-crabbe-pce-pce-initiated-lsp] specifies procedures that can be used for creation and deletion of PCE-initiated LSPs under the active stateful PCE model. However, this specification does not address GMPLS requirements outlined in the following:

- GMPLS support multiple switching capabilities on per TE link basis. GMPLS LSP creation requires knowledge of LSP switching capability (e.g., TDM, L2SC, OTN-TDM, LSC, etc.) to be used [RFC3471], [RFC3473].
- GMPLS LSP creation requires knowledge of the encoding type (e.g., lambda photonic, Ethernet, SONET/ SDH, G709 OTN, etc.) to be used by the LSP [RFC3471], [RFC3473].
- GMPLS LSP creation requires information of the generalized payload (G-PID) to be carried by the LSP [RFC3471], [RFC3473].

Internet-Draft      draft-ali-pce-remote-initiated-gmpls-lsp-01.txt

- GMPLS LSP creation requires specification of data flow specific traffic parameters (also known as Tspec), which are technology specific.
- GMPLS also specifies support for asymmetric bandwidth requests [RFC6387].
- GMPLS extends the addressing to include unnumbered interface identifiers, as defined in [RFC3477].
- In some technologies path calculation is tightly coupled with label selection along the route. For example, path calculation in a WDM network may include lambda continuity and/ or lambda feasibility constraints and hence a path computed by the PCE is associated with a specific lambda (label). Hence, in such networks, the label information needs to be provided to a PCC in order for a PCE to initiate GMPLS LSPs under the active stateful PCE model. I.e., explicit label control may be required.
- GMPLS specifies protection context for the LSP, as defined in [RFC4872] and [RFC4873].

#### 4. Remote Initiated LSP Usage Requirement

The requirement to specify usage of the LSP to the PCC includes but not limited to specification of the following information.

- The target IGP instance for the Remote-initiated LSP needs to be specified.
- In the target IGP instance, should the PCE-initiated LSP be advertised as a forwarding adjacency and/ or routing adjacency and/ or to be used locally by the PCC?
- Should the as Remote-initiated LSP be advertised an IPv4 FA/ RA, IPv6 FA/ RA or as unnumbered FA/ RA.
- If Remote-initiated LSP is to be advertised an IPv4 FA/ RA, IPv6 FA/ RA, what is the local and remote IP address is to be used for the advertisement.

#### 5. PCEP Extensions for Remote-Initiated GMPLS LSPs

Section 3 outlines GMPLS and application requirements that need to be satisfied in order for a PCE to initiate GMPLS LSPs under the active stateful PCE model. The section provides PCEP protocol extensions required to meet these requirements.

Internet-Draft      draft-ali-pce-remote-initiated-gmpls-lsp-01.txt

LSP create message defined in [I-D. draft-crabbe-pce-pce-initiated-lsp] needs to be extended to include GMPLS specific PCEP objects as follows:

### 5.1. Generalized Endpoint in LSP Create Message

This document does not modify the usage of END-POINTS object for PCE initiated LSPs as specified in [I-D. draft-crabbe-pce-pce-initiated-lsp]. It augments the usage as specified below.

END-POINTS object has been extended by [I-D. draft-ietf-pcep-gmpls-ext] to include a new object type called ''Generalized Endpoint''. PCCreate message sent by a PCE to a PCC to trigger a GMPLS LSP instantiation SHOULD include the END-POINTS with Generalized Endpoint object type. Furthermore, the END-POINTS object MUST contain ''label request'' TLV. The label request TLV is used to specify the switching type, encoding type and GPID of the LSP being instantiated by the PCE.

As mentioned earlier, the PCE server is assumed to be receiving topology data. In the use case of higher-layer signaling trigger, the addresses associated with bandwidth-on-demand interfaces are included, e.g., OTE-IP-R1, OTE-IP-R2 and OTE-IP-R3, in the use case example. These addresses and R1, R2 and R3 router IDs are used to derive source and destination address of the END-POINT object. As previously mentioned, in the case of NMS-VNMT cooperation model with L3 PCE, VNTM must receive such inter-layer interface association to configure the whole path.

The unnumbered endpoint TLV can be used to specify unnumbered endpoint addresses for the LSP being instantiated by the PCE. The END-POINTS MAY contain other TLVs defined in [I-D. draft-ietf-pcep-gmpls-ext].

If the END-POINTS Object of type Generalized Endpoint is missing the label request TLV, the PCC MUST send a PCErr message with Error-type=6 (Mandatory Object missing) and Error-value= TBA (LSP request TLV missing).

If the PCC does not support the END-POINTS Object of type Generalized Endpoint, the PCC MUST send a PCErr message with Error-type= ??? and Error-value= ???. [??? = already defined values to be looked up].

### 5.2. GENERALIZED-BANDWIDTH object in LSP Create Message

LSP create message defined in [I-D. draft-crabbe-pce-pce-initiated-lsp] can optionally include the BANDWIDTH object. However, the following possibilities cannot be represented in the BANDWIDTH object:

Internet-Draft      draft-ali-pce-remote-initiated-gmpls-lsp-01.txt

- Asymmetric bandwidth (different bandwidth in forward and reverse direction), as described in [RFC6387].

- Technology specific GMPLS parameters (e.g., Tspec for SDH/SONET, G.709, ATM, MEF, etc.) are not supported.

GENERALIZED-BANDWIDTH object has been defined in [I-D. draft-ietf-pcep-gmpls-ext] to address the above-mentioned limitation of the BANDWIDTH object.

This document specifies the use of GENERALIZED-BANDWIDTH object in PCCreate message. Specifically, GENERALIZED-BANDWIDTH object MAY be included in the PCCreate message. The GENERALIZED-BANDWIDTH object in PCCreate message is used to specify technology specific Tspec and asymmetrical bandwidth values for the LSP being instantiated by the PCE.

### 5.3. Protection Attributes in LSP Create Message

This document does not modify the usage of LSPA object for PCE initiated LSPs as specified in [I-D. draft-crabbe-pce-pce-initiated-lsp]. It augments the usage of LSPA object in LSP Create Message to carry the end-to-end protection context this also includes the protection state information.

The LSP Protection Information TLV of LSPA in the PCCreate message can be used to specify protection attributes of the LSP being instantiated by the PCE.

### 5.4. ERO in LSP Create Object

This document does not modify the usage of ERO object for PCE initiated LSPs as specified in [I-D. draft-crabbe-pce-pce-initiated-lsp]. It augments the usage as specified in the following sections.

#### 5.4.1. ERO with explicit label control

As mentioned earlier, there are technologies and scenarios where active stateful PCE requires explicit label control in order to instantiate an LSP.

Explicit label control (ELC) is a procedure supported by RSVP-TE, where the outgoing label(s) is (are) encoded in the ERO. [I-D. draft-ietf-pcep-gmpls-ext] extends the <ERO> object of PCEP to include explicit label control. The ELC procedure enables the PCE to provide such label(s) directly in the path ERO.

The extended ERO object in PCCreate message can be used to specify label along with ERO to PCC for the LSP being instantiated by the active stateful PCE.

#### 5.4.2. ERO with Path Keys

There are many scenarios in packet and optical networks where the route information of an LSP may not be provided to the PCC for confidentiality reasons. A multi-domain or multi-layer network is an example of such networks. Similarly, a GMPLS User-Network Interface (UNI) [RFC4208] is also an example of such networks.

In such scenarios, ERO containing the entire route cannot be provided to PCC (by PCE). Instead, PCE provides an ERO with Path Keys to the PCC. For example, in the case UNI interface between the router and the optical nodes, the ERO in the LSP Create Message may be constructed as follows:

- The first hop is a strict hop that provides the egress interface information at PCC. This interface information is used to get to a network node that can extend the rest of the ERO. (Please note that in the cases where the network node is not directly connected with the PCC, this part of ERO may consist of multiple hops and may be loose).
- The following(s) hop in the ERO may provide the network node with the path key [RFC5520] that can be resolved to get the contents of the route towards the destination.
- There may be further hops but these hops may also be encoded with the path keys (if needed).

This document does not change encoding or processing roles for the path keys, which are defined in [RFC5520].

#### 5.4.3. Switch Layer Object

[draft-ietf-pce-inter-layer-ext-07] specifies the SWITCH-LAYER object which defines and specifies the switching layer (or layers) in which a path MUST or MUST NOT be established. A switching layer is expressed as a switching type and encoding type. [I-D. draft-ietf-pcep-gmpls-ext], which defines the GMPLS extensions for PCEP, suggests using the SWITCH-LAYER object. Thus, SWITCH-LAYER object can be used in the PCCreate message to specify the switching layer (or layers) of the LSP being remotely initiated.

## 6. PCEP extension for PCEP Initiated LSP Usage Specification

The requirement to specify on how PCC should use the PCEP initiated LSPs is outlined in Section 4. This subsection specifies PCEP extension used to satisfy this requirement.

PCEP extensions specified in this section are equally applicable to PCEP initiated MPLS as well as GMPLS LSPs.

### 6.1. LSP\_TUNNEL\_INTERFACE\_ID Object in LSP Create Message

[RFC6107] defines LSP\_TUNNEL\_INTERFACE\_ID Object for communicating usage of the forwarding or routing adjacency from the ingress node to the egress node. This document extends the LSP Create Message to include LSP\_TUNNEL\_INTERFACE\_ID object defined in [RFC6107]. Object class and type for the LSP\_TUNNEL\_INTERFACE\_ID object are as follows:

Object Name: LSP\_TUNNEL\_INTERFACE\_ID

Object-Class Value: TBA by Iana (suggested value: 40)

Object-type: 1

The contents of this object are identical in encoding to the contents of the RSVP-TE LSP\_TUNNEL\_INTERFACE\_ID object defined in [RFC6107] and [RFC3477]. The following TLVs of RSVP-TE LSP\_TUNNEL\_INTERFACE\_ID object are acceptable in this object. The PCEP LSP\_TUNNEL\_INTERFACE\_ID object's TLV types correspond to RSVP-TE LSP\_TUNNEL\_INTERFACE\_ID object's TLV types. Please note that use of TLV type 1 defined in [RFC3477] is not specified by this document.

TLV Type	TLV Description	Reference
2	IPv4 interface identifier with target IGP instance	[RFC6107]
3	IPv6 interface identifier with target IGP instance	[RFC6107]
4	Unnumbered interface with target IGP instance	[RFC6107]

The meanings of the fields of PCEP LSP\_TUNNEL\_INTERFACE\_ID object are identical to those defined for the RSVP-TE LSP\_TUNNEL\_INTERFACE\_ID object. Similarly, meanings of the fields of PCEP LSP\_TUNNEL\_INTERFACE\_ID object's supported TLV are identical to those defined for the corresponding RSVP-TE LSP\_TUNNEL\_INTERFACE\_ID object's TLVs. The following fields have slightly different usage.

- IPv4 Interface Address field in IPv4 interface identifier with target IGP instance TLV: This field indicates the local IPv4 address to be assigned to the tunnel at the PCC (ingress node for RSVP-TE signaling). In the example use case of Section 2, IP address TUN-IP-R1 (tunnel IP address at R1) is carried in this field (if TUN-IP-R1 is a v4 address).
- IPv6 Interface Address field in IPv4 interface identifier with target IGP instance TLV: This field indicates the local IPv6 address to be assigned to the tunnel at the PCC (ingress node for RSVP-TE signaling). In the example use case of Section 2, IP address TUN-IP-R1 (tunnel IP address at R1) is carried in this field (if TUN-IP-R1 is a v6 address).
- LSR's Router ID field in Unnumbered interface with target IGP instance: The PCC SHOULD use the LSR's Router ID in Unnumbered interface with target IGP instance in advertising the LSP being initiated by the PCE. In the example use case of Section 2, this field carries router-id of R1 in the target IGP instance.
- Interface ID (32 bits) field in unnumbered interface with target IGP instance: All bits of this field MUST be set to 0 by the PCE server and MUST be ignored by PCC. PCC SHOULD allocate an Interface ID that fulfills Interface ID requirements specified in [RFC3477].

When the Ingress PCC receives an LPS Request Message with LSP\_TUNNEL\_INTERFACE\_ID TLV, it uses the information contained in the TLV to drive the IGP instance, treatment of the LSP being initiated in the target IGP instance (e.g., FA, RA or local usage), the local IPv4 or IPv6 address or router-id for unnumbered case to be used for advertisement of the LSP being instantiated.

## 6.2. Communicating LSP usage to Egress node

PCE does not need to send LSP Create message to egress node (node R2 in the example of section 2) to communicate LSP usage information. Instead PCC (Ingress signaling node) uses RSVP-TE signaling mechanism specified in [RFC6107] to send the LSP usage to Egress node. Specifically, when the Ingress PCC receives an LPS Request Message with LSP\_TUNNEL\_INTERFACE\_ID TLV, it SHOULD add LSP\_TUNNEL\_INTERFACE\_ID object in RSVP TE Path message. For this purpose, it is RECOMMENDED that the ingress PCC uses content of the LSP\_TUNNEL\_INTERFACE\_ID TLV in LSP Create Message in PCEP to drive LSP\_TUNNEL\_INTERFACE\_ID object in RSVP-TE. This document does not modify usage of LSP\_TUNNEL\_INTERFACE\_ID Object in RSVP-TE signaling as specified in [RFC6107].





Internet-Draft     draft-ali-pce-remote-initiated-gmpls-lsp-01.txt

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

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

[RFC5623] Oki, E., Takeda, T., Le Roux, JL., and A. Farrel, "Framework for PCE-Based Inter-Layer MPLS and GMPLS Traffic Engineering", RFC 5623, September 2009.

[RFC 6107] Shiomoto, K., Ed., and A. Farrel, Ed., "Procedures for Dynamically Signaled Hierarchical Label Switched Paths", RFC 6107, February 2011.

## 10.2. Informative References

[RFC3471] Berger, L., Ed., "Generalized Multi-Protocol Label Switching (GMPLS) Signaling Functional Description", RFC 3471, January 2003.

[RFC3473] Berger, L., Ed., "Generalized Multi-Protocol Label Switching (GMPLS) Signaling Resource ReserVation Protocol-Traffic Engineering (RSVP-TE) Extensions", RFC 3473, January 2003.

[RFC 5467] Berger, L., Takacs, A., Caviglia, D., Fedyk, D., and J. Meuric, "GMPLS Asymmetric Bandwidth Bidirectional Label Switched Paths (LSPs)", RFC 5467, March 2009.

[RFC3477] Kompella, K. and Y. Rekhter, "Signalling Unnumbered Links in Resource ReSerVation Protocol - Traffic Engineering (RSVP-TE)", RFC 3477, January 2003.

[RFC4872] Lang, J., Ed., Rekhter, Y., Ed., and D. Papadimitriou, Ed., "RSVP-TE Extensions in Support of End-to-End Generalized Multi-Protocol Label Switching (GMPLS) Recovery", RFC 4872, May 2007.

[RFC4873] Berger, L., Bryskin, I., Papadimitriou, D., and A. Farrel, "GMPLS Segment Recovery", RFC 4873, May 2007.

Internet-Draft      draft-ali-pce-remote-initiated-gmpls-lsp-01.txt

[RFC4208] Swallow, G., Drake, J., Ishimatsu, H., and Y. Rekhter,  
"Generalized Multiprotocol Label Switching (GMPLS)  
User-Network Interface (UNI): Resource ReserVation  
Protocol-Traffic Engineering (RSVP-TE) Support for the  
Overlay Model", RFC 4208, October 2005.

[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, April 2009.

#### Authors' Addresses

Zafar Ali  
Cisco Systems  
Email: zali@cisco.com

Siva Sivabalan  
Cisco Systems  
Email: msiva@cisco.com

Clarence Filsfils  
Cisco Systems  
Email: cfilsfil@cisco.com

Robert Varga  
Pantheon Technologies

Victor Lopez  
Telefonica I+D  
Email: vlopez@tid.es

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

PCE Working Group  
Internet-Draft  
Intended status: Informational  
Expires: January 10, 2014

R. Casellas, Ed.  
CTTC  
C. Margaria  
Coriant  
A. Farrel  
Old Dog Consulting  
O. Gonzalez de Dios  
Telefonica I+D  
D. Dhody  
X. Zhang  
Huawei Technologies  
July 09, 2013

Current issues with existing RBNF notation for PCEP messages and  
extensions  
draft-cmfg-pce-pcep-grammar-01

## Abstract

The PCEP protocol has been defined in [RFC5440] and later extended in several RFCs. This document aims at documenting inconsistencies when implementing a set of extensions and at providing a reference, complete and formal RBNF grammar for PCEP messages, including object ordering and precedence rules.

## Status of This Memo

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

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

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

This Internet-Draft will expire on January 10, 2014.

## Copyright Notice

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

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

## Table of Contents

1. Requirements Language . . . . .	2
2. Introduction and Motivation . . . . .	3
2.1. Object Ordering . . . . .	3
2.2. Inconsistent Naming . . . . .	5
2.3. Semantics and Exclusive Rules . . . . .	6
3. Initial Considerations . . . . .	8
4. RBNF Grammars . . . . .	8
4.1. Common Constructs . . . . .	8
4.1.1. Object Sequences . . . . .	8
4.1.2. Synchronized Vectors . . . . .	9
4.1.3. Monitoring Metrics . . . . .	9
4.1.4. Monitoring Requests and Responses . . . . .	10
4.2. PCEP Open Message . . . . .	10
4.3. PCEP Keep Alive (KeepAlive) Message . . . . .	10
4.4. PCEP Request (PCReq) Message . . . . .	11
4.5. PCEP Reply (PCRep) Message . . . . .	13
4.6. PCEP Monitoring Request (PCMonReq) Message . . . . .	13
4.7. PCEP Monitoring Reply (PCMonRep) Message . . . . .	13
4.8. PCEP Notify (PCNtf) Message . . . . .	14
4.9. PCEP Error (PCErr) Message . . . . .	14
4.10. PCEP Report (PCRpt) Message . . . . .	15
4.11. PCEP Update (PCUpd) Message . . . . .	15
5. Management Considerations . . . . .	15
6. Contributing Authors . . . . .	15
7. Acknowledgments . . . . .	15
8. Normative References . . . . .	15
Authors' Addresses . . . . .	17

## 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. Introduction and Motivation

The RBNF notation, defined in [RFC5511], is used to specify the message format for the Path Computation Element Communication Protocol (PCEP). The core of PCEP has been defined in [RFC5440] and later extended in [RFC5441], to support the Backward Recursive Path Computation (BRPC) procedure; in [RFC5455], adding a CLASSTYPE object to support Diffserv-aware Traffic Engineering (DS-TE); in [RFC5520], for topology confidentiality by means of Path keys; in [RFC5521], in support of exclusions; in [RFC5541] to convey specific objective functions; in [RFC5557], for Global Concurrent Optimization, in [RFC5886], for monitoring and in [RFC6006] for point-to-multipoint (P2MP) computation.

Most PCEP RFCs describe specific protocol extensions and, as such, they focus on their constructs extending some base RFCs. Although it is not the intention of each individual draft or RFC to provide the latest and most complete/full definition of the protocol messages, in practice combining all the extensions as defined in the respective RFCs is complex.

Message rules are sometimes provided within the text, resulting in ambiguity. Moreover, the fact that extensions may be defined in parallel may be a problem. The canonical example is the case where RFC X defines construct `p ::= A` and subsequent RFC Y extends RFC X stating that object C MUST follow object A and RFC Z also extends RFC X stating that object D MUST follow object A.

### 2.1. Object Ordering

The use of RBNF [RFC5511] states that the ordering of objects and constructs in an assignment is explicit, and protocol specifications MAY opt to state that ordering is only RECOMMENDED (the elements of a list of objects and constructs MAY be received in any order).

The core PCEP document [RFC5440] states in Section 6 that an implementation MUST form the PCEP messages using the object ordering specified in [RFC5440].

[RFC5886] equally states that "An implementation MUST form the PCEP messages using the object ordering specified in this document."

[RFC5521] only states that "the XRO is OPTIONAL and MAY be carried within Path Computation Request (PCReq) and Path Computation Reply

(PCRep) messages." and no ordering is provided. It does not mention SVEC objects or rules.

[RFC5541] specifies that "the OF object MAY be carried within a PCReq message. If an objective function is to be applied to a set of synchronized path computation requests, the OF object MUST be carried just after the corresponding SVEC (Synchronization VECtor) object and MUST NOT be repeated for each elementary request. Similarly, if a metric is to be applied to a set of synchronized requests, the METRIC object MUST follow the SVEC object and MUST NOT be repeated for each elementary request. (...) An OF object specifying an objective function that applies to an individual path computation request (non-synchronized case) MUST follow the RP object for which it applies". It should be understood that this last sentence must be relaxed or is in contradiction with the ENDPOINTS object.

RFCs that extend the core PCEP protocol are not consistent with the object ordering. For example, [RFC5520] defines:

```
<segment-computation> ::=
  <END-POINTS>
  [<LSPA>]
  [<BANDWIDTH>]
  [<BANDWIDTH>]
  [<metric-list>]
  (snip)
```

and states that "the format of the message for use in normal path computation is unmodified". However, [RFC5520] was not updated to reflect that the the BANDWIDTH object used for reoptimization was moved to appear after the RRO for which it applies, as given in [RFC5440] (updated in Errata ID: 3582):

```
<request> ::= <RP>
  <END-POINTS>
  [<LSPA>]
  [<BANDWIDTH>]
  [<metric-list>]
  [<RRO> [<BANDWIDTH>]]
  [<IRO>]
  [<LOAD-BALANCING>]
```

[RFC5541] in section 3.2 is not consistent with the ordering of OF and metric-list:

```
<svec-list> ::= <SVEC>
               [<OF>]
               [<metric-list>]

<request> ::= <RP>
               (snip)
               [<metric-list>]
               [<OF>]

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

In view of the above considerations, this document aims at providing an object ordering for PCEP messages so implementations can interoperate. Implementations conforming to this document **MUST** use the object ordering specified here.

## 2.2. Inconsistent Naming

PCEP RFCs may use inconsistent or ambiguous naming. For example [RFC5440] defines the Open message as having a common header and an OPEN object, and later uses Open to refer to the object that may appear in a PCErr message.

```
<Open Message> ::= <Common Header>
                   <OPEN>

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

It is common that a sequence or repetition of an object OBJ is noted as obj-list. It may happen that in extensions to core documents, the naming is kept although it no longer applies to such a sequence. For example, [RFC5886] states:



```
<svec-list> ::= <SVEC>
               [<OF>]
               [<svec-list>]
```

and later

```
<svec-list> ::= <SVEC>
               [<svec-list>]
```

### 2.3. Semantics and Exclusive Rules

The current RBNF notation does not capture the semantics/intent of the messages; notably, when two options are mutually exclusive and at least one is mandatory. In most cases, this is noted as both options being optional. For example [RFC5440] states:

```
<response>::=<RP>
             [<NO-PATH>]
             [<attribute-list>]
             [<path-list>]
```

with this example, a message that contains a response of the form <RP><NO-PATH><ERO><..> (that is, a NO-PATH object followed by a path) is correct and successfully parsed. Likewise, a response with just an RP object is valid. Although the actual text within the RFC may state the intention and disambiguate the grammar, having a RBNF notation that better captures semantics, message structure and original intent, enables the development of automated parsers that closely map the specification.

Similarly, if the intent is to specific a rule such as metric-pce which includes a PCE-ID object followed by a PROC-TIME object and/or an OVERLOAD object, the syntax:

```
<metric-pce> ::= <PCE-ID> [<PROC-TIME>] [<OVERLOAD>]
```

allows, amongst other combinations, that neither PROC-TIME nor OVERLOAD appears, which is not the intended behavior (there should be at least one metric). The alternative

```
<metric-pce> ::= <PCE-ID> <metric-argument-list>
<metric-argument-list> ::= <metric-argument> [<metric-argument-list>]
<metric-argument> ::= <PROC-TIME> | <OVERLOAD>
```

or equivalently

```
<metric-pce> ::= <PCE-ID> (<metric-argument>...)
<metric-argument> ::= <PROC-TIME> | <OVERLOAD>
```

does not reflect that each metric-argument should appear at most once. This can be addressed verbosely:

```
<metric-pce> ::= <PCE-ID>
                ( <PROC-TIME> | <OVERLOAD> | <PROC-TIME><OVERLOAD> )

<metric-pce> ::= <PCE-ID>
                ( <PROC-TIME>[<OVERLOAD>] | [<PROC-TIME>]<OVERLOAD> )
```

Here the semantic is that we require any object of the set {PROC-TIME, OVERLOAD} to be present, and there should be at least one. Note that currently there are only a few cases where the "non-empty set" case arises.

[Editor note/AF To make a normative or machine-readable definition, new notation could be defined:

```
- non-empty set, repetition not allowed
  <set> ::= { <a> | <b> | <c> }
- non-empty set, repetition allowed
  <set> ::= { <a> <b> <c> }
-- also can be expressed using the previous
   definition with
  <set> ::= { <a>... | <b>... | <c>... }
```

Note that the other options can already be handled

```
- non-repetition set allowed to be empty
  <set> ::= [<a>] [<b>] [<c>]
- repetition set allowed to be empty
  <set> ::= [<a>] [<b>] [<c>] [<set>]
```

The notation with "{" would be convenient to express implicit ordering (<a><a><b> ok but <a><b><a> not)].

A more condensed notation extension to the RBNF notation could also use a "sequential or" notation:

$\langle a \rangle \mid \mid \langle b \rangle$  is defined as  $\langle a \rangle \mid \langle b \rangle \mid \langle a \rangle \langle b \rangle$

$\langle a \rangle \mid \mid \langle b \rangle \mid \mid \langle c \rangle$  is defined as (assoc.)

$(\langle a \rangle \mid \langle b \rangle \mid \langle a \rangle \langle b \rangle) \mid \langle c \rangle \mid (\langle a \rangle \mid \langle b \rangle \mid \langle a \rangle \langle b \rangle) \langle c \rangle =$   
 $(\langle a \rangle \mid \langle b \rangle \mid \langle a \rangle \langle b \rangle) \mid \langle c \rangle \mid (\langle a \rangle \langle c \rangle \mid \langle b \rangle \langle c \rangle \mid \langle a \rangle \langle b \rangle \langle c \rangle) =$   
 $\langle a \rangle \mid \langle b \rangle \mid \langle c \rangle \mid \langle a \rangle \langle b \rangle \mid \langle a \rangle \langle c \rangle \mid \langle b \rangle \langle c \rangle \mid \langle a \rangle \langle b \rangle \langle c \rangle$

The use of sequential-or notation allows writing:

$\langle \text{metric-pce} \rangle ::= \langle \text{PCE-ID} \rangle ( \langle \text{PROC-TIME} \rangle \mid \mid \langle \text{OVERLOAD} \rangle )$

The goal of this document is then, first, to provide an (almost) formal (reasonably) complete definition of PCEP messages, checking the overall protocol and extensions consistency, defining an object ordering; and to set the basis for implementation agreements that aim at integrating published PCEP extensions. It is also a goal to provide alternative (although compatible) RBNF notations to be expressive enough to avoid invalid cases.

### 3. Initial Considerations

This document does not modify the content of defined PCEP objects and TLVs.

This document is not normative, the normative definition is included in the existing specs. This does not preclude integration with a future revision of such documents.

### 4. RBNF Grammars

This section provides the proposed RBNF notation for the PCEP messages. Specific constructs or grammar rules that appear in several messages or deserve special considerations are described first.

#### 4.1. Common Constructs

##### 4.1.1. Object Sequences

$\langle \text{of-list} \rangle \quad \quad \quad ::= \langle \text{OF} \rangle [ \langle \text{of-list} \rangle ]$   
 $\langle \text{metric-list} \rangle \quad \quad \quad ::= \langle \text{METRIC} \rangle [ \langle \text{metric-list} \rangle ]$   
 $\langle \text{vendor-info-list} \rangle \quad \quad \quad ::= \langle \text{VENDOR-INFORMATION} \rangle [ \langle \text{vendor-info-list} \rangle ]$

```
<pce-id-list> ::= <PCE-ID> [<pce-id-list>]  
-- (note: named pce-list in original)
```

#### 4.1.2. Synchronized Vectors

SVEC tuple:

A svec-tuple is a construct that associates a SVEC object with one or more constraining objects. The selected order follows the relative order of having OF and metric-list after the SVEC object, and the name svec-list has been changed since it no longer means a list of SVEC objects.

```
<svec-tuple> ::= <SVEC>  
                [<OF>]  
                [<metric-list>]  
                [<vendor-info-list>]  
                [<GC>]  
                [<XRO>]  
  
<svec-tuple-list> ::= <svec-tuple> [<svec-tuple-list>]
```

Note that [I-D.ietf-pce-vendor-constraints] defines:

```
<svec-list> ::= <SVEC>  
                [<OF>]  
                [<GC>]  
                [<XRO>]  
                [<metric-list>]  
                [<vendor-info-list>]  
                [<svec-list>]
```

The construct is updated to reflect the new name and to have the same relative order in the attributes that constrain a individual request

#### 4.1.3. Monitoring Metrics

A metric-pce-id is a rule that associates a PCE identified by its PCE-ID to a list of metric arguments.

```
<metric-pce-id> ::= <PCE-ID>  
                    (<PROC-TIME> [<OVERLOAD>] |  
                     [<PROC-TIME>] <OVERLOAD> )  
  
<metric-pce-id-list> ::= <metric-pce-id> [<metric-pce-id-list>]
```

#### 4.1.4. Monitoring Requests and Responses

See [RFC5886] for the definition of specific/general and in-band/out-of-band.

```
<monitoring> ::= <MONITORING> <PCC-ID-REQ>

<monitoring-request> ::= <monitoring> [<pce-id-list>]

<monitoring-response> ::= <monitoring>
    (<specific-monitoring-metrics-list> |
     <general-monitoring-metrics-list>)

<specific-monitoring-metrics-list> ::=
    <specific-monitoring-metrics>
    [<specific-monitoring-metrics-list>]

<general-monitoring-metrics-list> ::=
    <general-monitoring-metrics>
    [<general-monitoring-metrics-list>]

<specific-monitoring-metrics> ::=
    <RP> <monitoring-metrics>

<general-monitoring-metrics> ::=
    <monitoring-metrics>

<monitoring-metrics> ::=
    <metric-pce-id-list>
```

#### 4.2. PCEP Open Message

```
<Open Message> ::= <Common Header>
    <OPEN>
```

#### 4.3. PCEP Keep Alive (KeepAlive) Message

```
<KeepAlive Message> ::= <Common Header>
```

#### 4.4. PCEP Request (PCReq) Message

Note that the actual parsing depends on the content (flags) of the Request Parameters (RP) object, notably expansion and P2MP. In some cases, this may be considered redundant, e.g. the presence of a PATH\_KEY object and the corresponding flag.

[Editor's note: from a notation perspective, we lack a way to express "if object a field x has value v then include object b, else include object c". A possible way would be to define new intermediate types :

<a with x=v> and <a with x!=v> then

(<a with x=v> <b>) | (<a with x!=v> <c>)

this issue is still open.]

The PCReq message contains a possibly monitored list of requests, some of which may be grouped by SVEC tuples.

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

where:

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

-- A request is either an expansion, a P2P request or a P2MP request

```
<request>         ::= <expansion> |
                    <p2p_computation> |
                    <p2mp_computation>
```

```
<expansion>       ::= <RP><PATH-KEY>
```

```
<p2p_computation> ::= <RP><ENDPOINTS>
                    [<LSP>][<gen-bw>][<p2p-attributes>...]
```

```
<p2mp_computation> ::= <RP><tree-list>
                    [<p2mp-attributes>...]
```

-- For a P2P computation

```

<p2p-attributes> ::= <attributes>|<rro-bw-pair>

<attributes>      ::= <attribute> [<attributes>]

<attribute>       ::=
    <CLASSTYPE> |
    <LSPA> |
    <OF> |
    <metric-list> |
    <vendor-info-list> |
    <IRO> |
    <BNC> |
    <XRO> |
    <gen-load-balancing> |
    <INTER-LAYER> |
    <SWITCH-LAYER> |
    <REQ-ADAP-CAP>

-- in RFC6006 there is a bw per tree,
-- it is intended to be an optimization for an RRO list

<rro-bw-pair>     ::= <RRO> [<gen-bw>]

<rro-list-bw>     ::= (<RRO>...)[<gen-bw>]

<tree>            ::= <ENDPOINTS>(<rro-bw-pair>|<rro-list-bw>)

<gen-bw>          ::= <BANDWIDTH>[<GENERALIZED-BANDWIDTH>...]
-- per RFC5440 section 7.7

<gen-load-balancing> ::= <LOAD-BALANCING> |
                        <GENERALIZED-LOAD-BALANCING>

-- For P2MP computations - note some atts (BNC) are only P2MP

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

<tree> ::= <ENDPOINTS> <rro_bw_pair>

<p2mp-attributes> ::= (<attribute> | <BNC>) [<p2mp-attributes>]

```

## 4.5. PCEP Reply (PCRep) Message

```

<PCRep Message> ::= <Common Header>
                    [<svec-tuple-list>]
                    <response-list>

-- Note: should clarify the use of SVEC tuple list
where

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

-- An individual response may include monitoring info

<response> ::= <RP> [<monitoring>]
               (<success> | <failure>) [<monitoring-metrics>]

-- Note: should clarify P2MP attributes

<success> ::= <path-list>

<failure> ::= <NO-PATH> [<attributes>]

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

<path> ::= <ERO> <gen-bw> [<attributes>]

```

## 4.6. PCEP Monitoring Request (PCMonReq) Message

The PCMonReq message is defined in [RFC5886] for out-of-band monitoring requests.

[RFC5886] specifies that there is one mandatory object but the grammar also includes PCC-ID-REQ as mandatory.

[Ed note:does it make sense to include a pce-id-list and a svec-list/request-list at the same time?]

```

<PCMonReq Message> ::= <Common Header>
                       <monitoring-request>
                       [[<svec-tuple-list>] <request-list>]

```

## 4.7. PCEP Monitoring Reply (PCMonRep) Message



The PCMonRep message is defined in [RFC5886] for out-of-band monitoring responses.

[RFC5886] specifies that there is one mandatory object but the grammar also includes PCC-ID-REQ as mandatory.

[RFC5886] does not allow bundling several specific monitoring responses. A PCMonReq message causes N PCMonRep messages.

```
<PCMonRep Message> ::= <Common Header>
                        <monitoring-response>
```

#### 4.8. PCEP Notify (PCNtf) Message

```
<PCNtf Message> ::= <Common Header>
                    ( <solicited-notify> | <unsolicited-notify> )
```

where

```
<solicited-notify>   ::= <request-id-list> <notification-list>
```

```
<unsolicited-notify> ::= <notification-list>
```

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

```
<notification-list> ::= <NOTIFICATION> [<notification-list>]
```

#### 4.9. PCEP Error (PCErr) Message

Errors can occur during PCEP handshake, or bound to one or more requests.

An error during handshake is never solicited, i.e., not associated to a list of requests.

A solicited error binds one or more Requests (RPs) to one or more PCEP-ERROR objects.

```
<PCErr Message> ::= <Common Header>
                    ( <solicited-error> | <unsolicited-error> )
```

where

```
<solicited-error>   ::= <request-id-list> <pcep-error-list>
```

```
<unsolicited-error> ::= <handshake-error> | <pcep-error-list>
```

```
<handshake-error> ::= <pcep-error-list> <OPEN>

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

<pcep-error-list> ::= <PCEP-ERROR> [<pcep-error-list>]
```

#### 4.10. PCEP Report (PCRpt) Message

TBD see [I-D.ietf-pce-stateful-pce].

#### 4.11. PCEP Update (PCUpd) Message

TBD see [I-D.ietf-pce-stateful-pce].

### 5. Management Considerations

TBD

### 6. Contributing Authors

Robert Varga  
Pantheon  
robert.varga@pantheon.sk

### 7. Acknowledgments

TBD

### 8. Normative References

- [I-D.ietf-pce-stateful-pce]  
Crabbe, E., Medved, J., Minei, I., and R. Varga, "PCEP Extensions for Stateful PCE", draft-ietf-pce-stateful-pce-05 (work in progress), July 2013.
- [I-D.ietf-pce-vendor-constraints]  
Zhang, F. and A. Farrel, "Conveying Vendor-Specific Constraints in the Path Computation Element Protocol", draft-ietf-pce-vendor-constraints-10 (work in progress), April 2013.
- [I-D.ietf-pce-wson-rwa-ext]  
Lee, Y. and R. Casellas, "PCEP Extension for WSON Routing and Wavelength Assignment", draft-ietf-pce-wson-rwa-ext-00 (work in progress), April 2013.

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC4655] Farrel, A., Vasseur, J., and J. Ash, "A Path Computation Element (PCE)-Based Architecture", RFC 4655, August 2006.
- [RFC5440] Vasseur, JP. and JL. Le Roux, "Path Computation Element (PCE) Communication Protocol (PCEP)", RFC 5440, March 2009.
- [RFC5441] Vasseur, JP., Zhang, R., Bitar, N., and JL. Le Roux, "A Backward-Recursive PCE-Based Computation (BRPC) Procedure to Compute Shortest Constrained Inter-Domain Traffic Engineering Label Switched Paths", RFC 5441, April 2009.
- [RFC5455] Sivabalan, S., Parker, J., Boutros, S., and K. Kumaki, "Diffserv-Aware Class-Type Object for the Path Computation Element Communication Protocol", RFC 5455, March 2009.
- [RFC5511] Farrel, A., "Routing Backus-Naur Form (RBNF): A Syntax Used to Form Encoding Rules in Various Routing Protocol Specifications", RFC 5511, April 2009.
- [RFC5520] Bradford, R., Vasseur, JP., and A. Farrel, "Preserving Topology Confidentiality in Inter-Domain Path Computation Using a Path-Key-Based Mechanism", RFC 5520, April 2009.
- [RFC5521] Oki, E., Takeda, T., and A. Farrel, "Extensions to the Path Computation Element Communication Protocol (PCEP) for Route Exclusions", RFC 5521, April 2009.
- [RFC5541] Le Roux, JL., Vasseur, JP., and Y. Lee, "Encoding of Objective Functions in the Path Computation Element Communication Protocol (PCEP)", RFC 5541, June 2009.
- [RFC5557] Lee, Y., Le Roux, JL., King, D., and E. Oki, "Path Computation Element Communication Protocol (PCEP) Requirements and Protocol Extensions in Support of Global Concurrent Optimization", RFC 5557, July 2009.
- [RFC5886] Vasseur, JP., Le Roux, JL., and Y. Ikejiri, "A Set of Monitoring Tools for Path Computation Element (PCE)-Based Architecture", RFC 5886, June 2010.
- [RFC5886] Vasseur, JP., Le Roux, JL., and Y. Ikejiri, "A Set of Monitoring Tools for Path Computation Element (PCE)-Based Architecture", RFC 5886, June 2010.

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

#### Authors' Addresses

Ramon Casellas (editor)  
CTTC  
Av. Carl Friedrich Gauss n.7  
Castelldefels 08860 Barcelona  
Spain

Phone: +34 93 645 29 00  
Email: ramon.casellas@cttc.es

Cyril Margaria  
Coriant  
St.-Martin-Str. 76  
Muenchen 81541  
Germany

Phone: +49 89 5159 16934  
Email: cyril.margaria@coriant.com

Adrian Farrel  
Old Dog Consulting  
  
Email: adrian@olddog.co.uk

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

Phone: +34913128832  
Email: ogondio@tid.es

Dhruv Dhody  
Huawei Technologies  
Leela Palace  
Bangalore, Karnataka 560008  
INDIA

Email: [dhruv.dhody@huawei.com](mailto:dhruv.dhody@huawei.com)

Xian Zhang  
Huawei Technologies

Email: [zhang.xian@huawei.com](mailto:zhang.xian@huawei.com)

Network Working Group  
Internet-Draft  
Intended status: Standards Track  
Expires: January 13, 2014

E. Crabbe  
Google, Inc.  
I. Minei  
Juniper Networks, Inc.  
S. Sivabalan  
Cisco Systems, Inc.  
R. Varga  
Pantheon Technologies SRO  
July 12, 2013

PCEP Extensions for PCE-initiated LSP Setup in a Stateful PCE Model  
draft-crabbe-pce-pce-initiated-lsp-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 extensions described in [I-D.ietf-pce-stateful-pce] provide stateful control of Multiprotocol Label Switching (MPLS) Traffic Engineering Label Switched Paths (TE LSP) via PCEP, for a model where the PCC delegates control over one or more locally configured LSPs to the PCE. This document describes the creation and deletion of PCE-initiated LSPs under the stateful PCE model.

Requirements Language

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

Status of this Memo

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

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

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

This Internet-Draft will expire on January 13, 2014.

#### Copyright Notice

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

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

## Table of Contents

1. Introduction . . . . .	4
2. Terminology . . . . .	4
3. Architectural Overview . . . . .	4
3.1. Motivation . . . . .	4
3.2. Operation overview . . . . .	5
4. Support of PCE-initiated LSPs . . . . .	6
4.1. Stateful PCE Capability TLV . . . . .	6
5. PCE-initiated LSP instantiation and deletion . . . . .	7
5.1. The LSP Initiate Message . . . . .	7
5.2. The R flag in the SRP Object . . . . .	8
5.3. LSP instantiation . . . . .	9
5.4. LSP deletion . . . . .	10
6. LSP delegation and cleanup . . . . .	10
7. IANA considerations . . . . .	11
7.1. PCEP Messages . . . . .	11
7.2. PCEP-Error Object . . . . .	11
8. Security Considerations . . . . .	12
8.1. Malicious PCE . . . . .	12
9. Acknowledgements . . . . .	12
10. References . . . . .	13
10.1. Normative References . . . . .	13
10.2. Informative References . . . . .	14
Authors' Addresses . . . . .	14



## 1. Introduction

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

Stateful pce [I-D.ietf-pce-stateful-pce] specifies a set of extensions to PCEP to enable stateful control of TE LSPs between and across PCEP sessions in compliance with [RFC4657]. It includes mechanisms to effect LSP state synchronization between PCCs and PCEs, delegation of control of LSPs to PCEs, and PCE control of timing and sequence of path computations within and across PCEP sessions and focuses on a model where LSPs are configured on the PCC and control over them is delegated to the PCE.

This document 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.

## 2. Terminology

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

This document uses the following terms defined in [I-D.ietf-pce-stateful-pce]: Stateful PCE, Delegation, Redelegation Timeout, State Timeout Interval LSP State Report, LSP Update Request.

The following terms are defined in this document:

PCE-initiated LSP: LSP that is instantiated as a result of a request from the PCE.

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

## 3. Architectural Overview

### 3.1. Motivation

[I-D.ietf-pce-stateful-pce] provides stateful control over LSPs that are locally configured on the PCC. This model relies on the LER taking an active role in delegating locally configured LSPs to the

PCE, and is well suited in environments where the LSP placement is fairly static. However, in environments where the LSP placement needs to change in response to application demands, it is useful to support dynamic creation and tear down of LSPs. The ability for a PCE to trigger the creation of LSPs on demand can make possible agile software-driven network operation, and can be seamlessly integrated into a controller-based network architecture, where intelligence in the controller can determine when and where to set up paths.

A possible use case is one of a software-driven network, where applications request network resources and paths from the network infrastructure. For example, an application can request a path with certain constraints between two LSRs by contacting the PCE. The PCE can compute a path satisfying the constraints, and instruct the head end LSR to instantiate and signal it. When the path is no longer required by the application, the PCE can request its teardown.

Another use case is one of dynamically adjusting aggregate bandwidth between two points in the network using multiple LSPs. This functionality is very similar to auto-bandwidth, but allows for providing the desired capacity through multiple LSPs. This approach overcomes two of the limitations auto-bandwidth can experience: 1) growing the capacity between the endpoints beyond the capacity of individual links in the path and 2) achieving good bin-packing through use of several small LSPs instead of a single large one. The number of LSPs varies based on the demand, and LSPs are created and deleted dynamically to satisfy the bandwidth requirements.

Another use case is that of demand engineering, where a PCE with visibility into both the network state and the demand matrix can anticipate and optimize how traffic is distributed across the infrastructure. Such optimizations may require creating new paths across the infrastructure.

### 3.2. Operation overview

A PCC indicates its ability to support PCE provisioned dynamic LSPs during the PCEP Initialization Phase via a new flag in the STATEFUL-PCE-CAPABILITY TLV (see details in Section 4.1).

The decision when to instantiate or delete a PCE-initiated LSP is out of the scope of this document. To instantiate or delete an LSP, the PCE sends a new message, the Path Computation LSP Initiate Request (PCInitiate) message to the PCC. The LSP Initiate Request MUST include the SRP and LSP objects, and the LSP object MUST include the Symbolic Path Name TLV.

For an instantiation operation, the PCE MUST include the ERO and END-

POINTS object and may include various attributes as per [RFC5440]. The PCC creates the LSP using the attributes communicated by the PCE, and local values for the unspecified parameters. It assigns a unique PLSP-ID for the LSP and automatically delegates the LSP to the PCE. It also generates an LSP State Report (PCRpt) for the LSP, carrying the newly assigned PLSP-ID and indicating the delegation via the delegation bit. The PCE may update the attributes of the LSP via subsequent PCUpd messages. Subsequent LSP State Report and LSP Update Request for the LSP will carry the PCC-assigned PLSP-ID, which uniquely identifies the LSP. See details in Section 5.3.

Once instantiated, the delegation procedures for PCE-initiated LSPs are the same as for PCC initiated LSPs. This applies to the case of a PCE failure as well. In order to allow for network cleanup without manual intervention, the PCC SHOULD support removal of PCE-initiated LSPs as one of the behaviors applied on expiration of the State Timeout Interval [I-D.ietf-pce-stateful-pce]. The behavior SHOULD be picked based on local policy, and can result either in LSP removal, or into reverting to operator-defined default parameters. See details in Section 6.

To indicate a delete operation, the PCE MUST use the R flag in the SRP object. As a result of the deletion request, the PCC MUST remove all state related to the LSP, and send a PCRpt with the R flag set for the removed state. See details in Section 5.3.

#### 4. Support of PCE-initiated LSPs

A PCC indicates its ability to support PCE provisioned dynamic LSPs during the PCEP Initialization phase. The Open Object in the Open message contains the "Stateful PCE Capability" TLV, defined in [I-D.ietf-pce-stateful-pce]. A new flag, the I (LSP-INSTANTIATION-CAPABILITY) flag is introduced to indicate support for instantiation of PCE-initiated LSPs. A PCE can initiate LSPs only for PCCs that advertised this capability and a PCC will follow the procedures described in this document only on sessions where the PCE advertised the I flag.

##### 4.1. Stateful PCE Capability TLV

The format of the STATEFUL-PCE-CAPABILITY TLV is shown in the following figure:

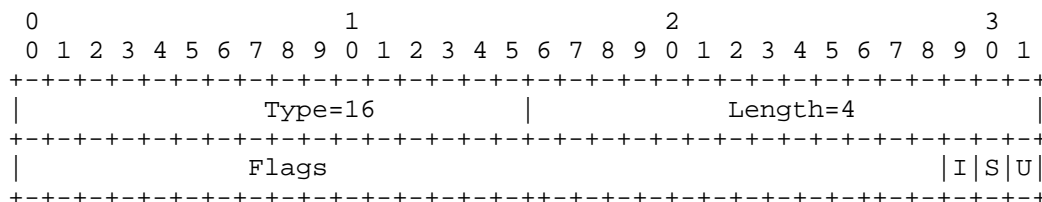


Figure 1: STATEFUL-PCE-CAPABILITY TLV format

The type of the TLV is defined in [I-D.ietf-pce-stateful-pce] and it has a fixed length of 4 octets.

The value comprises a single field - Flags (32 bits). The U and S bits are defined in [I-D.ietf-pce-stateful-pce].

I (LSP-INSTANTIATION-CAPABILITY - 1 bit): If set to 1 by a PCC, the I Flag indicates that the PCC allows instantiation of an LSP by a PCE. If set to 1 by a PCE, the I flag indicates that the PCE will attempt to instantiate LSPs. The LSP-INSTANTIATION-CAPABILITY flag must be set by both PCC and PCE in order to support PCE-initiated LSP instantiation.

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

## 5. PCE-initiated LSP instantiation and deletion

To initiate an LSP, a PCE sends a PCInitiate message to a PCC. The message format, objects and TLVs are discussed separately below for the creation and the deletion cases.

### 5.1. The LSP Initiate Message

A Path Computation LSP Initiate Message (also referred to as PCInitiate message) is a PCEP message sent by a PCE to a PCC to trigger LSP instantiation or deletion. The Message-Type field of the PCEP common header for the PCInitiate message is set to [TBD]. The PCInitiate message MUST include the SRP and the LSP objects, and may contain other objects, as discussed later in this section. If either the SRP or the LSP object is missing, the PCC MUST send a PCErr as described in [I-D.ietf-pce-stateful-pce]. LSP instantiation is done by sending an LSP Initiate Message with an LSP object with the reserved PLSP-ID 0. LSP deletion is done by sending an LSP Initiate Message with an LSP object carrying the PLSP-ID of the LSP to be removed and an SRP object with the R flag set (see Section 5.2).

The format of a PCInitiate message for LSP instantiation is as follows:

```
<PCInitiate Message> ::= <Common Header>
                           <PCE-initiated-lsp-list>
```

Where:

```
<PCE-initiated-lsp-list> ::= <PCE-initiated-lsp-request>[<PCE-initiated-lsp-list>]
```

```
<PCE-initiated-lsp-request> ::= <SRP>
                                <LSP>
                                <END-POINTS>
                                <ERO>
                                [<attribute-list>]
```

Where:

<attribute-list> is defined in [RFC5440] and extended by PCEP extensions.

The SRP object is used to correlate between initiation requests sent by the PCE and the error reports and state reports sent by the PCC. Every request from the PCE receives a new SRP-ID-number. This number is unique per PCEP session and is incremented each time an operation (initiation, update, etc) is requested from the PCE. The value of the SRP-ID-number MUST be echoed back by the PCC in PCErr and PCRpt messages to allow for correlation between requests made by the PCE and errors or state reports generated by the PCC. Details of the SRP object and its use can be found in [I-D.ietf-pce-stateful-pce].

## 5.2. The R flag in the SRP Object

The format of the SRP object is shown Figure 2:

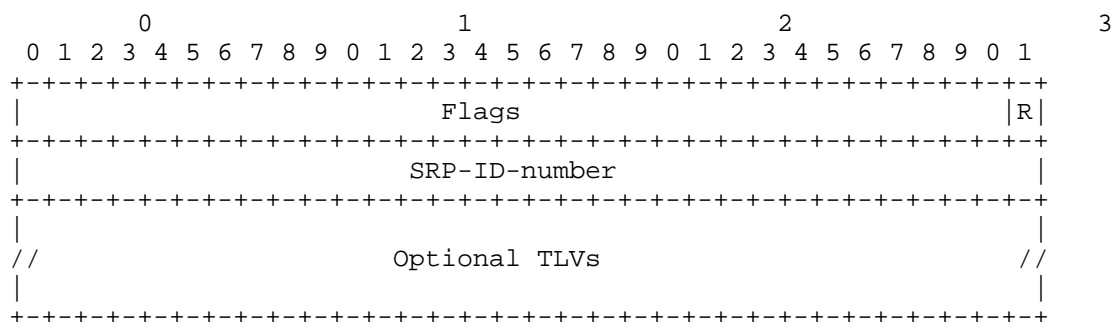


Figure 2: The SRP Object format

The type object is defined in [I-D.ietf-pce-stateful-pce].

A new flag is defined to indicate a delete operation initiated by the PCE:

R (LSP-REMOVE - 1 bit): If set to 1, it indicates a removal request initiated by the PCE.

### 5.3. LSP instantiation

LSP instantiation is done by sending an LSP Initiate Message with an LSP object with the reserved PLSP-ID 0.

Receipt of a PCInitiate Message with a non-zero PLSP-ID and the R flag in the SRP object set to zero results in a PCErr message of type 19 (Invalid Operation) and value 8 (non-zero PLSP-ID in LSP initiation request).

The LSP-sig-type field in the LSP is set to the signaling type that is requested for the LSP setup. This draft defines procedures for RSVP-signaled LSPs.

The END-POINTS Object is mandatory for an instantiation request of an RSVP-signaled LSP. It contains the source and destination addresses for provisioning the LSP. If the END-POINTS Object is missing, the PCC MUST send a PCErr message with Error-type=6 (Mandatory Object missing) and Error-value=3 (END-POINTS Object missing).

The ERO Object is mandatory for an instantiation request. It contains the ERO for the LSP. If the ERO Object is missing, the PCC MUST send a PCErr message with Error-type=6 (Mandatory Object missing) and Error-value=9 (ERO Object missing).

The LSP Object MUST include the SYMBOLIC-PATH-NAME TLV, which will be used to correlate between the PCC-assigned PLSP-ID and the LSP. If the TLV is missing, the PCC MUST send a PCErr message with Error-type=6 (Mandatory object missing) and Error-value=14 (SYMBOLIC-PATH-NAME TLV missing). The symbolic name used for provisioning PCE-initiated LSPs must not have conflict with the LSP name of any existing LSP in the PCC. (Existing LSPs may be either statically configured, or initiated by another PCE). If there is conflict with the LSP name, the PCC MUST send a PCErr message with Error-type=23 (Bad Parameter value) and Error-value=1 (SYMBOLIC-PATH-NAME in use). The only exception to this rule is for LSPs for which the State timeout timer is running (see Section 6).

The PCE MAY include various attributes as per [RFC5440]. The PCC MUST use these values in the LSP instantiation, and local values for

unspecified parameters. After the LSP setup, the PCC MUST send a PCRpt to the PCE, reflecting these values. The SRP object in the PCRpt message MUST echo the value of the PCInitiate message that triggered the setup.

If the PCC determines that the LSP parameters proposed in the PCInitiate message are unacceptable, it MUST trigger a PCErr with error-type=TBD (PCE instantiation error) and error-value=1 (Unacceptable instantiation parameters). If the PCC encounters an internal error during the processing of the PCInitiate message, it MUST trigger a PCErr with error-type=TBD (PCE instantiation error) and error-value=2 (Internal error).

A PCC MUST relay to the PCE errors it encounters in the setup of PCE-initiated LSP by sending a PCErr with error-type=TBD (PCE instantiation error) and error-value=3 (RSVP signaling error). The PCErr MUST echo the SRP-id-number of the PCInitiate message. The PCEP-ERROR object SHOULD include the RSVP Error Spec TLV (if an ERROR SPEC was returned to the PCC by a downstream node).

A PCC SHOULD be able to place a limit on either the number of LSPs or the percentage of resources that are allocated to honor PCE-initiated LSP requests. As soon as that limit is reached, the PCC MUST send a PCErr message of type 19 (Invalid Operation) and value TBD "PCE-initiated limit reached" and is free to drop any incoming PCInitiate messages without additional processing.

#### 5.4. LSP deletion

PCE-initiated removal of a PCE-initiated LSP is done by setting the R (remove) flag in the SRP Object in the PCInitiate message from the PCE. The LSP is identified by the PLSP-ID in the LSP object. If the PLSP-ID is unknown, the PCC MUST generate a PCErr with error type 19, error value 3, "Unknown PLSP-ID". A PLSP-ID of zero removes all LSPs that were initiated by the PCE. If the PLSP-ID specified in the PCInitiate message is not delegated to the PCE, the PCC MUST send a PCErr message indicating "LSP is not delegated" (Error code 19, error value 1 ([I-D.ietf-pce-stateful-pce])). Following the removal of the LSP, the PCC MUST send a PCRpt as described in [I-D.ietf-pce-stateful-pce].

#### 6. LSP delegation and cleanup

PCE-initiated LSPs are automatically delegated by the PCC to the PCE upon instantiation. The PCC MUST delegate the LSP to the PCE by setting the delegation bit to 1 in the PCRpt that includes the assigned PLSP-ID. All subsequent messages from the PCC must have the

delegation bit set to 1. The PCC cannot revoke the delegation for PCE-initiated LSPs for an active PCEP session. Sending a PCRpt message with the delegation bit set to 0 results in a PCErr message of type 19 (Invalid Operation) and value TBD "Delegation for PCE-initiated LSP cannot be revoked".

A PCE MAY return a delegation to the PCC, to allow for LSP transfer between PCEs. Doing so MUST trigger the State Timeout Interval timer ([I-D.ietf-pce-stateful-pce]).

In case of PCEP session failure, control over PCE-initiated LSPs reverts to the PCC at the expiration of the redelegation timeout. To obtain control of a PCE-initiated LSP, a PCE (either the original or one of its backups) sends a PCInitiate message, including just the SRP and LSP objects, and carrying the PLSP-ID of the LSP it wants to take control of. Receipt of a PCInitiate message with a non-zero PLSP-ID normally results in the generation of a PCErr. If the State Timeout timer is running, the PCC MUST NOT generate an error and redelegate the LSP to the PCE. The State Timeout timer is stopped upon the redelegation.

The State Timeout timer ensures that a PCE crash does not result in automatic and immediate disruption for the services using PCE-initiated LSPs. PCE-initiated LSPs are not be removed immediately upon PCE failure. Instead, they are cleaned up on the expiration of this timer. This allows for network cleanup without manual intervention. The PCC SHOULD support removal of PCE-initiated LSPs as one of the behaviors applied on expiration of the State Timeout Interval [I-D.ietf-pce-stateful-pce]. The behavior SHOULD be picked based on local policy, and can result either in LSP removal, or into reverting to operator-defined default parameters.

## 7. IANA considerations

### 7.1. PCEP Messages

This document defines the following new PCEP messages:

Value	Meaning	Reference
12	Initiate	This document

### 7.2. PCEP-Error Object

This document defines new Error-Type and Error-Value for the following new error conditions:



Error-Type	Meaning
6	Mandatory Object missing Error-value=13: LSP cleanup TLV missing Error-value=14: SYMBOLIC-PATH-NAME TLV missing
19	Invalid operation Error-value=6: PCE-initiated LSP limit reached Error-value=7: Delegation for PCE-initiated LSP cannot be revoked Error-value=8: Non-zero PLSP-ID in LSP initiation request
23	Bad parameter value Error-value=1: SYMBOLIC-PATH-NAME in use
24	LSP instantiation error Error-value=1: Unacceptable instantiation parameters Error-value=2: Internal error Error-value=3: RSVP signaling error

## 8. Security Considerations

The security considerations described in [I-D.ietf-pce-stateful-pce] apply to the extensions described in this document. Additional considerations related to a malicious PCE are introduced.

### 8.1. Malicious PCE

The LSP instantiation mechanism described in this document allows a PCE to generate state on the PCC and throughout the network. As a result, it introduces a new attack vector: an attacker may flood the PCC with LSP instantiation requests and consume network and LSR resources, either by spoofing messages or by compromising the PCE itself.

A PCC can protect itself from such an attack by imposing a limit on either the number of LSPs or the percentage of resources that are allocated to honor PCE-initiated LSP requests. As soon as that limit is reached, the PCC MUST send a PCErr message of type 19 (Invalid Operation) and value 3 "PCE-initiated LSP limit reached" and is free to drop any incoming PCInitiate messages without additional processing.

Rapid flaps triggered by the PCE can also be an attack vector. This will be discussed in a future version of this document.

## 9. Acknowledgements

We would like to thank Jan Medved, Ambrose Kwong and Raveendra Trovi

for their contributions to this document.

## 10. References

### 10.1. Normative References

- [I-D.ietf-pce-stateful-pce]  
Crabbe, E., Medved, J., Minei, I., and R. Varga, "PCEP Extensions for Stateful PCE", draft-ietf-pce-stateful-pce-05 (work in progress), July 2013.
- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC2205] Braden, B., Zhang, L., Berson, S., Herzog, S., and S. Jamin, "Resource ReSerVation Protocol (RSVP) -- Version 1 Functional Specification", RFC 2205, September 1997.
- [RFC3209] Awduche, D., Berger, L., Gan, D., Li, T., Srinivasan, V., and G. Swallow, "RSVP-TE: Extensions to RSVP for LSP Tunnels", RFC 3209, December 2001.
- [RFC4090] Pan, P., Swallow, G., and A. Atlas, "Fast Reroute Extensions to RSVP-TE for LSP Tunnels", RFC 4090, May 2005.
- [RFC5088] Le Roux, JL., Vasseur, JP., Ikejiri, Y., and R. Zhang, "OSPF Protocol Extensions for Path Computation Element (PCE) Discovery", RFC 5088, January 2008.
- [RFC5089] Le Roux, JL., Vasseur, JP., Ikejiri, Y., and R. Zhang, "IS-IS Protocol Extensions for Path Computation Element (PCE) Discovery", RFC 5089, January 2008.
- [RFC5226] Narten, T. and H. Alvestrand, "Guidelines for Writing an IANA Considerations Section in RFCs", BCP 26, RFC 5226, May 2008.
- [RFC5440] Vasseur, JP. and JL. Le Roux, "Path Computation Element (PCE) Communication Protocol (PCEP)", RFC 5440, March 2009.
- [RFC5511] Farrel, A., "Routing Backus-Naur Form (RBNF): A Syntax Used to Form Encoding Rules in Various Routing Protocol Specifications", RFC 5511, April 2009.

## 10.2. Informative References

- [RFC2702] Awduche, D., Malcolm, J., Agogbua, J., O'Dell, M., and J. McManus, "Requirements for Traffic Engineering Over MPLS", RFC 2702, September 1999.
- [RFC3031] Rosen, E., Viswanathan, A., and R. Callon, "Multiprotocol Label Switching Architecture", RFC 3031, January 2001.
- [RFC3346] Boyle, J., Gill, V., Hannan, A., Cooper, D., Awduche, D., Christian, B., and W. Lai, "Applicability Statement for Traffic Engineering with MPLS", RFC 3346, August 2002.
- [RFC3630] Katz, D., Kompella, K., and D. Yeung, "Traffic Engineering (TE) Extensions to OSPF Version 2", RFC 3630, September 2003.
- [RFC4655] Farrel, A., Vasseur, J., and J. Ash, "A Path Computation Element (PCE)-Based Architecture", RFC 4655, August 2006.
- [RFC4657] Ash, J. and J. Le Roux, "Path Computation Element (PCE) Communication Protocol Generic Requirements", RFC 4657, September 2006.
- [RFC5305] Li, T. and H. Smit, "IS-IS Extensions for Traffic Engineering", RFC 5305, October 2008.
- [RFC5394] Bryskin, I., Papadimitriou, D., Berger, L., and J. Ash, "Policy-Enabled Path Computation Framework", RFC 5394, December 2008.
- [RFC5557] Lee, Y., Le Roux, J.L., King, D., and E. Oki, "Path Computation Element Communication Protocol (PCEP) Requirements and Protocol Extensions in Support of Global Concurrent Optimization", RFC 5557, July 2009.

## Authors' Addresses

Edward Crabbe  
Google, Inc.  
1600 Amphitheatre Parkway  
Mountain View, CA 94043  
US

Email: edc@google.com

Ina Minei  
Juniper Networks, Inc.  
1194 N. Mathilda Ave.  
Sunnyvale, CA 94089  
US

Email: ina@juniper.net

Siva Sivabalan  
Cisco Systems, Inc.  
170 West Tasman Dr.  
San Jose, CA 95134  
US

Email: msiva@cisco.com

Robert Varga  
Pantheon Technologies SRO  
Mlynske Nivy 56  
Bratislava 821 05  
Slovakia

Email: robert.varga@pantheon.sk



PCE Working Group  
Internet-Draft  
Intended status: Standards Track  
Expires: January 5, 2014

D. Dhody  
U. Pallé  
Q. Zhao  
Huawei Technology  
D. King  
Old Dog Consulting  
July 4, 2013

Management Information Base (MIB) for the PCE Communications Protocol  
(PCEP) for Path-Key based Confidentiality in Inter-Domain Path  
Computation.

draft-dhody-pce-pcep-pathkey-mib-05

## Abstract

This memo defines an experimental portion of the Management Information Base (MIB) for use with network management protocols in the Internet community. In particular, it describes managed objects for modeling of the Path Computation Element communication Protocol (PCEP) for communications between a Path Computation Client (PCC) and a Path Computation Element (PCE), or between two PCEs when path-key-based confidentiality in inter-domain path computation is requested.

## Status of This Memo

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

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

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

This Internet-Draft will expire on January 5, 2014.

## Copyright Notice

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

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents

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

## Table of Contents

1. Introduction . . . . .	3
1.1. Requirements Language . . . . .	3
2. Terminology . . . . .	3
3. The Internet-Standard Management Framework . . . . .	4
4. PCEP Pathkey MIB Module Architecture . . . . .	4
4.1. Relations to other MIB modules . . . . .	4
5. Example of the PCEP PathKey MIB module usage . . . . .	5
6. Object definitions . . . . .	6
6.1. PCEP-PATHKEY-MIB . . . . .	6
7. IANA Considerations . . . . .	20
8. Security Considerations . . . . .	20
9. References . . . . .	20
9.1. Normative References . . . . .	20
9.2. Informative References . . . . .	21

## 1. Introduction

The Path Computation Element (PCE) defined in [RFC4655] is an entity that is capable of computing a network path or route based on a network graph, and applying computational constraints. A Path Computation Client (PCC) may make requests to a PCE for paths to be computed.

The PCE communication protocol (PCEP) is designed as a communication protocol between PCCs and PCEs for path computations and is defined in [RFC5440].

If confidentiality is required between domains, Path-Key-Based mechanism is described in [RFC5520]. For preserving the confidentiality of the "Confidential Path Segment (CPS)"; the PCE returns a path containing a loose hop in place of the segment that must be kept confidential.

[PCEP-MIB] defines a portion of the MIB for use with network management protocols in the Internet community that can be used to manage PCEP communications between a PCC and a PCE, or between two PCEs. This memo describes MIB for path-key-based confidentiality in inter-domain path computations.

### 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

This document uses the terminology defined in [RFC4655], [RFC5440] and [RFC5520]. The following terminology is used in this document.

**Domain:** Any collection of network elements within a common sphere of address management or path computational responsibility. Examples of domains include Interior Gateway Protocol (IGP) areas and Autonomous Systems (ASs).

**IGP:** Interior Gateway Protocol. Either of the two routing protocols, Open Shortest Path First (OSPF) or Intermediate System to Intermediate System (IS-IS).



### 3. The Internet-Standard Management Framework

For a detailed overview of the documents that describe the current Internet-Standard Management Framework, please refer to section 7 of [RFC3410].

Managed objects are accessed via a virtual information store, termed the Management Information Base or MIB. MIB objects are generally accessed through the Simple Network Management Protocol (SNMP). Objects in the MIB are defined using the mechanisms defined in the Structure of Management Information (SMI). This memo specifies a MIB module that is compliant to the SMIV2, which is described in STD 58, RFC 2578 [RFC2578] and STD 58, RFC 2580 [RFC2580].

### 4. PCEP Pathkey MIB Module Architecture

The PCEP Pathkey MIB will contain the following information:

- o PCEP Pathkey counters, timers and configurations
- o PCEP Pathkey table of Confidential Path Segment (CPS) related information.

The PCEP Pathkey MIB has no role when PCEP peer is PCC.

#### 4.1. Relations to other MIB modules

The PCEP Pathkey MIB imports the following textual conventions from the MPLS-TC-STD-MIB defined in [RFC3811]:

- o MplsPathIndex
- o TeHopAddressType
- o TeHopAddress
- o TeHopAddressUnnum

The PCEP Pathkey MIB imports the following textual conventions from the INET-ADDRESS-MIB defined in [RFC4001]:

- o InetAddressType
- o InetAddress

## 5. Example of the PCEP PathKey MIB module usage

In this section we provide an example to showcase the relationship between `pcePcepPathKeyTable` and `pcePcepPathKeyHopTable` described in Section 6. While this example is not meant to illustrate every permutation of the MIB, nor in its entirety, it is intended as an aid to understand some of the key concepts. It is meant to be read after going through the MIB itself.

`pcePcepPathKeyTable` of the PCEP-PATHKEY-MIB module:

```
{
    pcePcepPathKey                (4512),
    pcePcepPathKeyCPSIndex        (1),
    pcePcepPathKeyReqSrcAddrType  ipv4 (1),
    pcePcepPathKeyReqSrcAddr      (1.1.1.1),
    pcePcepPathKeyRequestId       (10),
    pcePcepPathKeyRetrieved       (1),
    pcePcepPathKeyRtrAddrType     ipv4 (1),
    pcePcepPathKeyRtrAddr         (2.2.2.2),
    pcePcepPathKeyDiscardTime     (10),
    pcePcepPathKeyReuseTime       (30)
}
```

Entries of `pcePcepPathKeyHopTable` of the PCEP-PATHKEY-MIB module:

```
{
    pcePcepPathKeyHopListIndex    1,
    pcePcepPathKeyHopIndex        1,
    pcePcepPathKeyHopAddrType     ipv4 (1),
    pcePcepPathKeyHopIpAddr       "192.168.100.1",
    pcePcepPathKeyHopIpPrefixLen  32,
    pcePcepPathKeyHopAddrUnnum    0,
}
{
    pcePcepPathKeyHopListIndex    1,
    pcePcepPathKeyHopIndex        2,
    pcePcepPathKeyHopAddrType     ipv4 (1),
    pcePcepPathKeyHopIpAddr       "192.168.100.2",
    pcePcepPathKeyHopIpPrefixLen  32,
    pcePcepPathKeyHopAddrUnnum    0
}
```

The `pcePcepPathKeyTable` is the table for all the Path-Keys generated by PCE. To access the CPS hidden by path-key `pcePcepPathKey` (4512), index `pcePcepPathKeyCPSIndex` (1) is used in `pcePcepPathKeyHopTable` to find the hop list (`pcePcepPathKeyHopListIndex`). To access each hop of the path another index `pcePcepPathKeyHopIndex` is used along with `pcePcepPathKeyHopListIndex`.

## 6. Object definitions

## 6.1. PCEP-PATHKEY-MIB

```
PCEP-PATHKEY-MIB DEFINITIONS ::= BEGIN
```

```
IMPORTS
```

```
    MODULE-IDENTITY,
    OBJECT-TYPE,
    mib-2,
    NOTIFICATION-TYPE,
    Unsigned32,
    Counter32
        FROM SNMPv2-SMI                -- RFC 2578
    TruthValue,
    TimeStamp
        FROM SNMPv2-TC                -- RFC 2579
    MODULE-COMPLIANCE,
    OBJECT-GROUP,
    NOTIFICATION-GROUP
        FROM SNMPv2-CONF              -- RFC 2580
    MplsPathIndex,
    TeHopAddressType,
    TeHopAddress,
    TeHopAddressUnnum
        FROM MPLS-TC-STD-MIB          -- RFC 3811
    InetAddressType,
    InetAddress
        FROM INET-ADDRESS-MIB         -- RFC 4001
```

```
pcePcepPathkeyMIB MODULE-IDENTITY
```

```
    LAST-UPDATED
```

```
        "201307031200Z" -- July 03, 2013
```

```
    ORGANIZATION
```

```
        "IETF Path Computation Element (PCE) Working Group"
```

```
    CONTACT-INFO
```

```
        "Email: pce@ietf.org
```

```
        WG charter
```

```
        http://www.ietf.org/html.charters/pce-charter.html"
```

```
DESCRIPTION
```

```
"This MIB module defines a collection of objects for managing PCE
communication protocol(PCEP) for Path-Key-Based Inter-Domain Path
Computation"
```

```
-- Revision history
```

```
    REVISION
```

"201307031200Z" -- 03 July 2013 12:00:00 EST  
DESCRIPTION

"

Main Changes from -04 draft :

1. Aligment with the updates in PCEP-MIB draft
2. Editorial Changes.

REVISION

"201208171200Z" -- 17 Aug 2012 12:00:00 EST  
DESCRIPTION

"

Main Changes from -03 draft :

1. Adding of DEFVAL for some objects.
2. Editorial Changes.

REVISION

"201202221200Z" -- 22 Feb 2012 12:00:00 EST  
DESCRIPTION

"

Main Changes from -02 draft :

1. Editorial Changes.
2. Updated Contact Information.

REVISION

"201109051200Z" -- 05 Sept 2011 12:00:00 EST  
DESCRIPTION

"

Main Changes from -01 draft :

1. Added pcePcepPathKeyCPSIndex.
2. Added pcePcepPathKeyHopListIndex.
3. Removed pcePcepPathKeyHopNum.
4. Updated Contact Information.

REVISION

"201103081200Z" -- 08 Mar 2011 12:00:00 EST  
DESCRIPTION

"

Main Changes from -00 draft :

1. Added HopTable to store the CPS hops.
2. Added Path Key Creation Time.

REVISION

"201009171200Z" -- 17 Sep 2010 12:00:00 EST  
DESCRIPTION

"draft-00 version"

::= { experimental 9999 } --

```
pcePcepPathKeyNotifications OBJECT IDENTIFIER ::=
    { pcePcepPathkeyMIB 0 }
pcePcepPathKeyMIBObjects OBJECT IDENTIFIER ::=
    { pcePcepPathkeyMIB 1 }
pcePcepPathKeyConformance OBJECT IDENTIFIER ::=
    { pcePcepPathkeyMIB 2 }
pcePcepPathKeyObjects OBJECT IDENTIFIER ::=
    { pcePcepPathKeyMIBObjects 1 }

--

-- PCE Pathkey Objects

--

pcePcepPathKeyDiscardTimer OBJECT-TYPE
    SYNTAX  Unsigned32
    UNITS   "minutes"
    MAX-ACCESS read-only
    STATUS  mandatory
    DESCRIPTION
        "The value which indicates a period of time after the
        expiration of which a PCE can discard unwanted
        path-keys and CPS."
    DEFVAL {10}
    ::= { pcePcepPathKeyObjects 1 }

pcePcepPathKeyReUseTimer OBJECT-TYPE
    SYNTAX  Unsigned32
    UNITS   "minutes"
    MAX-ACCESS read-only
    STATUS  mandatory
    DESCRIPTION
        "The value which indicates a period of time which
        should expire before an old path-key could be
        reused for a new CPS."
    DEFVAL {30}
    ::= { pcePcepPathKeyObjects 2 }
```

```
pcePcepPathKeyRetainStatus OBJECT-TYPE
    SYNTAX      INTEGER {
                    enabled(1),
                    disabled(2)
                }
    MAX-ACCESS   read-only
    STATUS       optional
    DESCRIPTION
        "The path-key retain status of this PCE to retain the
        path-key and CPS after retrieval."
    DEFVAL {disabled(2)}
    ::= { pcePcepPathKeyObjects 3 }

pcePcepPathKeysGenerated OBJECT-TYPE
    SYNTAX      Counter32
    MAX-ACCESS   read-only
    STATUS       mandatory
    DESCRIPTION
        "The number of path-keys generated by this PCE."
    ::= { pcePcepPathKeyObjects 4 }

pcePcepPathKeyExpandUn OBJECT-TYPE
    SYNTAX      Counter32
    MAX-ACCESS   read-only
    STATUS       mandatory
    DESCRIPTION
        "The number of attempts to expand an unknown
        path-key."
    ::= { pcePcepPathKeyObjects 5 }

pcePcepPathKeyExpandExp OBJECT-TYPE
    SYNTAX      Counter32
    MAX-ACCESS   read-only
    STATUS       mandatory
    DESCRIPTION
        "The number of attempts to expand an expired
        path-key."
    ::= { pcePcepPathKeyObjects 6 }

pcePcepPathKeyExpandSame OBJECT-TYPE
    SYNTAX      Counter32
    MAX-ACCESS   read-only
    STATUS       optional
    DESCRIPTION
        "The number of attempts to expand the same
        path-key."
    ::= { pcePcepPathKeyObjects 7 }
```

```
pcePcepPathKeyExpNoExpansion OBJECT-TYPE
    SYNTAX Counter32
    MAX-ACCESS read-only
    STATUS optional
    DESCRIPTION
        "The number of path-keys expired without any attempt
        to expand it."
    ::= { pcePcepPathKeyObjects 8 }

pcePcepPathKeyExpansionSuccess OBJECT-TYPE
    SYNTAX Counter32
    MAX-ACCESS read-only
    STATUS optional
    DESCRIPTION
        "The number of path-key expansion requests (PCReq)
        which had successful retrieval."
    ::= { pcePcepPathKeyObjects 9 }

pcePcepPathKeyExpansionFailures OBJECT-TYPE
    SYNTAX Counter32
    MAX-ACCESS read-only
    STATUS optional
    DESCRIPTION
        "The number of path-key expansion requests (PCReq)
        which had failed retrieval."
    ::= { pcePcepPathKeyObjects 10 }

pcePcepPathKeyConfig OBJECT-TYPE
    SYNTAX INTEGER {
        enabled(1),
        disabled(2)
    }
    MAX-ACCESS read-only
    STATUS mandatory
    DESCRIPTION
        "Path-key based confidentiality is enabled."
    DEFVAL {disabled(2)}
    ::= { pcePcepPathKeyObjects 11 }

pcePcepPathKeyTable OBJECT-TYPE
    SYNTAX SEQUENCE OF pcePcepPathKeyEntry
    MAX-ACCESS not-accessible
    STATUS current
    DESCRIPTION
        "This table contains information about the
        Pathkey CPS of PCE. Applicable only when
        pcePcepPathKeyConfig is enabled(1)."
```

```
    ::= { pcePcepPathKeyObjects 12 }
```

```

pcePcepPathKeyEntry OBJECT-TYPE
    SYNTAX      pcePcepPathKeyEntry
    MAX-ACCESS  not-accessible
    STATUS      current
    DESCRIPTION
        "An entry in this table represents a path-key and
        CPS. An entry is only created when a path-key is
        generated by PCE during inter-domain path
        computation."

    INDEX      { pcePcepPathKey }

    ::= { pcePcepPathKeyTable 1 }

pcePcepPathKeyEntry ::= SEQUENCE {
    pcePcepPathKey                Unsigned32,
    pcePcepPathKeyCPSIndex        MplsPathIndex,
    pcePcepPathKeyReqSrcAddrType  InetAddressType,
    pcePcepPathKeyReqSrcAddr      InetAddress,
    pcePcepPathKeyRequestId       Unsigned32,
    pcePcepPathKeyRetrieved        INTEGER,
    pcePcepPathKeyRtrAddrType     InetAddressType,
    pcePcepPathKeyRtrAddr         InetAddress,
    pcePcepPathKeyCreationTime     TimeStamp,
    pcePcepPathKeyDiscardTime      Unsigned32,
    pcePcepPathKeyReuseTime        Unsigned32,
}

pcePcepPathKey OBJECT-TYPE
    SYNTAX      Unsigned32 (1..65535)
    MAX-ACCESS  read-only
    STATUS      mandatory
    DESCRIPTION
        "The path-key value to identify a CPS."
    ::= { pcePcepPathKeyEntry 1 }

pcePcepPathKeyCPSIndex OBJECT-TYPE
    SYNTAX      MplsPathIndex
    MAX-ACCESS  read-only
    STATUS      mandatory
    DESCRIPTION
        "The HopList index of the CPS. This index
        is used to expand Hops in
        pcePcepPathKeyHopTable."
    ::= { pcePcepPathKeyEntry 2 }

```



```
pcePcepPathKeyReqSrcAddrType OBJECT-TYPE
    SYNTAX      InetAddressType
    MAX-ACCESS  read-only
    STATUS      mandatory
    DESCRIPTION
        "The type of the PCEP peer Internet address.
        This object specifies how the value of the
        pcePcepPathKeyReqSrcAddr object should be
        interpreted."
    ::= { pcePcepPathKeyEntry 3 }

pcePcepPathKeyReqSrcAddr OBJECT-TYPE
    SYNTAX      InetAddress
    MAX-ACCESS  read-only
    STATUS      mandatory
    DESCRIPTION
        "The Internet address of the PCEP peer that
        issued the original request that led to the
        creation of the path-key.
        The type is given by
        pcePcepPathKeyReqSrcAddrType "
    ::= { pcePcepPathKeyEntry 4 }

pcePcepPathKeyRequestId OBJECT-TYPE
    SYNTAX      Unsigned32
    MAX-ACCESS  read-only
    STATUS      mandatory
    DESCRIPTION
        "The request ID of the original PCReq that led
        to the creation of the path-key."
    ::= { pcePcepPathKeyEntry 5 }

pcePcepPathKeyRetrieved OBJECT-TYPE
    SYNTAX      INTEGER {
                    TRUE(1),
                    FALSE(2)
                }
    MAX-ACCESS  read-only
    STATUS      mandatory
    DESCRIPTION
        "It specifies whether the path-key is retrieved
        or not."
    ::= { pcePcepPathKeyEntry 6 }
```

```
pcePcepPathKeyRtrAddrType OBJECT-TYPE
    SYNTAX  InetAddressType
    MAX-ACCESS read-only
    STATUS  mandatory
    DESCRIPTION
        "The type of the PCEP peer Internet address.
        This object specifies how the value of the
        pcePcepPathKeyRtrAddr object should be
        interpreted. Applicable only when
        pcePcepPathKeyRetrieved is TRUE(1)."
```

```
 ::= { pcePcepPathKeyEntry 7 }
```

```
pcePcepPathKeyRtrAddr OBJECT-TYPE
    SYNTAX  InetAddress
    MAX-ACCESS read-only
    STATUS  mandatory
    DESCRIPTION
        "The Internet address of the PCEP peer that
        issued the path-key expansion or retrieval.
        Applicable only when pcePcepPathKeyRetrieved
        is TRUE(1). The type is given by
        pcePcepPathKeyRtrAddrType."
```

```
 ::= { pcePcepPathKeyEntry 8 }
```

```
pcePcepPathKeyCreationTime OBJECT-TYPE
    SYNTAX  TimeStamp
    MAX-ACCESS read-only
    STATUS  mandatory
    DESCRIPTION
        "The value of sysUpTime at which Path Key
        was generated by PCE."
```

```
 ::= { pcePcepPathKeyEntry 9 }
```

```
pcePcepPathKeyDiscardTime OBJECT-TYPE
    SYNTAX  Unsigned32
    UNIT "seconds"
    MAX-ACCESS read-only
    STATUS  mandatory
    DESCRIPTION
        "The time after which the path segment associated
        with the path-key will be discarded."
```

```
 ::= { pcePcepPathKeyEntry 10 }
```

```

pcePcepPathKeyReuseTime OBJECT-TYPE
    SYNTAX      Unsigned32
    UNIT        "seconds"
    MAX-ACCESS  read-only
    STATUS      mandatory
    DESCRIPTION
        "The time after which the path-key will be available
        for re-use."
    ::= { pcePcepPathKeyEntry 11 }

pcePcepPathKeyHopTable OBJECT-TYPE
    SYNTAX      SEQUENCE OF pcePcepPathKeyHopEntry
    MAX-ACCESS  not-accessible
    STATUS      current
    DESCRIPTION
        "This table contains information about the
        Pathkey Hop in the CPS of PCE."
    ::= { pcePcepPathKeyObjects 12 }

pcePcepPathKeyHopEntry OBJECT-TYPE
    SYNTAX      pcePcepPathKeyHopEntry
    MAX-ACCESS  not-accessible
    STATUS      current
    DESCRIPTION
        "An entry in this table represents a Hop in the CPS.
        An entry is only created when a path-key generated
        by PCE during inter-domain computation."
    INDEX       { pcePcepPathKeyHopListIndex,
                  pcePcepPathKeyHopIndex }
    ::= { pcePcepPathKeyHopTable 1 }

pcePcepPathKeyHopEntry ::= SEQUENCE {
    pcePcepPathKeyHopListIndex      MplsPathIndex,
    pcePcepPathKeyHopIndex          MplsPathIndex,
    pcePcepPathKeyHopAddrType       TeHopAddressType,
    pcePcepPathKeyHopIpAddress      TeHopAddress,
    pcePcepPathKeyHopIpPrefixLen    InetAddressPrefixLength,
    pcePcepPathKeyHopAddrUnnum      TeHopAddressUnnum,
}

```

```
pcePcepPathKeyHopListIndex OBJECT-TYPE
    SYNTAX  MplsPathIndex
    MAX-ACCESS read-only
    STATUS  mandatory
    DESCRIPTION
        "The primary index into this table identifying a
        particular CPS. All hops in the CPS will have the
        same ListIndex. This corresponds to
        pcePcepPathKeyCPSIndex in pcePcepPathKeyEntry."

    ::= {  pcePcepPathKeyHopEntry 1  }

pcePcepPathKeyHopIndex OBJECT-TYPE
    SYNTAX  MplsPathIndex
    MAX-ACCESS read-only
    STATUS  mandatory
    DESCRIPTION
        "The secondry index into this table identifying a
        particular Hop in the CPS."

    ::= {  pcePcepPathKeyHopEntry 2  }

pcePcepPathKeyHopAddrType OBJECT-TYPE
    SYNTAX  TeHopAddressType
    MAX-ACCESS read-only
    STATUS  mandatory
    DESCRIPTION
        "The Hop Address Type of this CPS hop. Only
        ipv4(1), ipv6(2) and unnum(4) are allowed."
    DEFVAL { ipv4 }
    ::= {  pcePcepPathKeyHopEntry 3  }

pcePcepPathKeyHopIpAddr OBJECT-TYPE
    SYNTAX  TeHopAddress
    MAX-ACCESS read-only
    STATUS  mandatory
    DESCRIPTION
        "The Hop Address for this CPS hop.
        The type of this address is determined by the
        value of the corresponding
        pcePcepPathKeyHopAddrType."
    DEFVAL { '00000000'h } -- IPv4 address 0.0.0.0
    ::= {  pcePcepPathKeyHopEntry 4  }
```

```

pcePcepPathKeyHopIpPrefixLen OBJECT-TYPE
    SYNTAX InetAddressPrefixLength
    MAX-ACCESS read-only
    STATUS current
    DESCRIPTION
        "pcePcepPathKeyHopAddrType if set to ipv4(1) or
        ipv6(2), then this value will contain an
        appropriate prefix length for the IP address in
        object pcePcepPathKeyHopIpAddr. Otherwise this
        value is irrelevant and should be ignored."
    DEFVAL { 32 }
    ::= { pcePcepPathKeyHopEntry 5 }

pcePcepPathKeyHopAddrUnnum OBJECT-TYPE
    SYNTAX TeHopAddressUnnum
    MAX-ACCESS read-only
    STATUS current
    DESCRIPTION
        "If pcePcepPathKeyHopAddrType is set to unnum(4),
        then this value will contain the interface
        identifier of the unnumbered interface for this
        hop. This object should be used in conjunction
        with pcePcepPathKeyHopIpAddr which would contain
        the LSR Router ID in this case."
    ::= { pcePcepPathKeyHopEntry 6 }

---

--- Notifications

---

pcePcepPathKeyExpandUnNtf NOTIFICATION-TYPE
    OBJECTS {
        pcePcepPathKeyExpandUn
    }
    STATUS mandatory
    DESCRIPTION
        "This notification is sent when an attempt to expand
        an unknown path-key is made. The value of the
        counter pcePcepPathKeyExpandUn is also increased at
        this time."
    ::= { pcePcepPathKeyNotifications 1 }

```

```

pcePcepPathKeyExpandExpNtf NOTIFICATION-TYPE
    OBJECTS      {
                    pcePcepPathKeyExpandExp
                }
    STATUS        mandatory
    DESCRIPTION   "This notification is sent when an attempt to expand
                    an expired path-key is made. The value of the
                    counter pcePcepPathKeyExpandExp is also increased
                    at this time."
    ::= { pcePcepPathKeyNotifications 2 }

```

```

pcePcepPathKeyExpandSameNtf NOTIFICATION-TYPE
    OBJECTS      {
                    pcePcepPathKeyExpandSame
                }
    STATUS        optional
    DESCRIPTION   "This notification is sent when a duplicate attempt
                    to expand the same path-key is made. The value of
                    the counter pcePcepPathKeyExpandSame is also
                    increased at this time."
    ::= { pcePcepPathKeyNotifications 3 }

```

```

pcePcepPathKeyExpNoExpansionNtf NOTIFICATION-TYPE
    OBJECTS      {
                    pcePcepPathKeyExpNoExpansion
                }
    STATUS        optional
    DESCRIPTION   "This notification is sent when path-key expires
                    without any attempt to expand it. The value of
                    the counter pcePcepPathKeyExpNoExpansion is also
                    increased at this time."
    ::= { pcePcepPathKeyNotifications 4 }

```

```

--*****
-- Module Conformance Statement
--*****

```

```

pcePcepPathKeyGroups
    OBJECT IDENTIFIER ::= { pcePcepPathKeyConformance 1 }

```

```

pcePcepPathKeyCompliances
    OBJECT IDENTIFIER ::= { pcePcepPathKeyConformance 2 }

```

```
--
-- Read-Only Compliance
--

pcePcepPathKeyModuleReadOnlyCompliance MODULE-COMPLIANCE
    STATUS current
    DESCRIPTION
        "The Module is implemented with support
        for read-only.  In other words, only monitoring
        is available by implementing this
        MODULE-COMPLIANCE."

    MODULE -- this module
        MANDATORY-GROUPS
            { pcePcepPathKeyGeneralGroup,
              pcePcepPathKeyNotificationsGroup
            }
        ::= { pcePcepPathKeyCompliances 1 }

-- units of conformance
```

```
pcePcepPathKeyGeneralGroup OBJECT-GROUP
    OBJECTS {
        pcePcepPathKeyDiscardTimer,
        pcePcepPathKeyReUseTimer,
        pcePcepPathKeysGenerated,
        pcePcepPathKeyExpandUn,
        pcePcepPathKeyExpandExp,
        pcePcepPathKeyConfig,
        pcePcepPathKey,
        pcePcepPathKeyCPSIndex,
        pcePcepPathKeyReqSrcAddrType,
        pcePcepPathKeyReqSrcAddr,
        pcePcepPathKeyRequestId,
        pcePcepPathKeyRetrieved,
        pcePcepPathKeyRtrAddrType,
        pcePcepPathKeyRtrAddr,
        pcePcepPathKeyCreationTime,
        pcePcepPathKeyDiscardTime,
        pcePcepPathKeyReuseTime,
        pcePcepPathKeyHopListIndex,
        pcePcepPathKeyHopIndex,
        pcePcepPathKeyHopAddrType,
        pcePcepPathKeyHopIpAddr,
        pcePcepPathKeyHopIpPrefixLen,
        pcePcepPathKeyHopAddrUnnum,
    }
    STATUS      current
    DESCRIPTION
        "Objects that apply to all PCEP Pathkey MIB
        implementations."

    ::= { pcePcepPathKeyGroups 1 }

pcePcepPathKeyNotificationsGroup NOTIFICATION-GROUP
    NOTIFICATIONS { pcePcepPathKeyExpandUnNtf,
                    pcePcepPathKeyExpandExpNtf
    }
    STATUS      current
    DESCRIPTION
        "The notifications for a PCEP Pathkey MIB
        implementation."
    ::= { pcePcepPathKeyGroups 2 }

END
```



## 7. IANA Considerations

The MIB module in this document uses the following IANA-assigned OBJECT IDENTIFIER values recorded in the SMI Numbers registry:

Descriptor	OBJECT IDENTIFIER value
-----	-----
pcePcepPathkeyMIB	{ mib-2 XXX }

Editor's Note (to be removed prior to publication): the IANA is requested to assign a value for "XXX" under the 'mib-2' subtree and to record the assignment in the SMI Numbers registry. When the assignment has been made, the RFC Editor is asked to replace "XXX" (here and in the MIB module) with the assigned value and to remove this note.

## 8. Security Considerations

[PCEP-MIB] describes the security consideration related to the PCE MIB module, which are applicable to PCE Path-Key MIB defined in this document. Further [RFC5520] describes various security consideration when dealing with Path-Key. Since this MIB contains confidential path segment, care should be taken to maintain the confidentiality during SNMP MIB operations.

## 9. References

### 9.1. Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC2578] McCloghrie, K., Ed., Perkins, D., Ed., and J. Schoenwaelder, Ed., "Structure of Management Information Version 2 (SMIv2)", STD 58, RFC 2578, April 1999.
- [RFC2579] McCloghrie, K., Ed., Perkins, D., Ed., and J. Schoenwaelder, Ed., "Textual Conventions for SMIv2", STD 58, RFC 2579, April 1999.
- [RFC2580] McCloghrie, K., Perkins, D., and J. Schoenwaelder, "Conformance Statements for SMIv2", STD 58, RFC 2580, April 1999.
- [RFC2863] McCloghrie, K. and F. Kastenholz, "The Interfaces Group MIB", RFC 2863, June 2000.

- [RFC3411] Harrington, D., Presuhn, R., and B. Wijnen, "An Architecture for Describing Simple Network Management Protocol (SNMP) Management Frameworks", STD 62, RFC 3411, December 2002.
- [RFC3811] Nadeau, T. and J. Cucchiara, "Definitions of Textual Conventions (TCs) for Multiprotocol Label Switching (MPLS) Management", RFC 3811, June 2004.
- [RFC3813] Srinivasan, C., Viswanathan, A., and T. Nadeau, "Multiprotocol Label Switching (MPLS) Label Switching Router (LSR) Management Information Base (MIB)", RFC 3813, June 2004.
- [RFC4001] Daniele, M., Haberman, B., Routhier, S., and J. Schoenwaelder, "Textual Conventions for Internet Network Addresses", RFC 4001, February 2005.
- [RFC5440] Vasseur, JP. and JL. Le Roux, "Path Computation Element (PCE) Communication Protocol (PCEP)", RFC 5440, March 2009.

## 9.2. Informative References

- [RFC3410] Case, J., Mundy, R., Partain, D., and B. Stewart, "Introduction and Applicability Statements for Internet-Standard Management Framework", RFC 3410, December 2002.
- [RFC4655] Farrel, A., Vasseur, J., and J. Ash, "A Path Computation Element (PCE)-Based Architecture", RFC 4655, August 2006.
- [RFC5520] Bradford, R., Vasseur, JP., and A. Farrel, "Preserving Topology Confidentiality in Inter-Domain Path Computation Using a Path-Key-Based Mechanism", RFC 5520, April 2009.
- [PCEP-MIB] Kiran Koushik, A S., Emile, S., Zhao, Q., King, D., and J. Hardwick, "PCE communication protocol(PCEP) Management Information Base (draft-ietf-pce-pcep-mib-04)", February 2013.

Authors' Addresses

Dhruv Dhody  
Huawei Technology  
Leela Palace  
Bangalore, Karnataka 560008  
INDIA

EMail: dhruv.dhody@huawei.com

Udayasree Palle  
Huawei Technology  
Leela Palace  
Bangalore, Karnataka 560008  
INDIA

EMail: udayasree.palle@huawei.com

Quintin Zhao  
Huawei Technology  
125 Nagog Technology Park  
Acton, MA 01719  
US

EMail: quintin.zhao@huawei.com

Daniel King  
Old Dog Consulting  
UK

EMail: daniel@olddog.co.uk

PCE Working Group  
Internet-Draft  
Intended status: Standards Track  
Expires: January 5, 2014

D. Dhody  
U. Palle  
Huawei Technologies India Pvt  
Ltd  
July 4, 2013

PCEP Extensions for MPLS-TE LSP Automatic Bandwidth Adjustment with  
stateful PCE  
draft-dhody-pce-stateful-pce-auto-bandwidth-01

## Abstract

The Path Computation Element Communication Protocol (PCEP) provides mechanisms for Path Computation Elements (PCEs) to perform path computations in response to Path Computation Clients (PCCs) requests. The extensions described in [STATEFUL-PCE] provide stateful control of Multiprotocol Label Switching (MPLS) Traffic Engineering Label Switched Paths (TE LSP) via PCEP, for a model where the PCC delegates control over one or more locally configured LSPs to the PCE.

This document describes the automatic bandwidth adjustment of such LSPs under the stateful PCE model.

## Status of This Memo

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

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

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

This Internet-Draft will expire on January 5, 2014.

## Copyright Notice

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

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents

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

## Table of Contents

1. Introduction . . . . .	3
1.1. Requirements Language . . . . .	3
2. Terminology . . . . .	4
3. Architectural Overview . . . . .	4
3.1. Auto-Bandwidth Overview . . . . .	4
3.2. Deploying Auto-Bandwidth Feature . . . . .	5
4. Extensions to the PCEP . . . . .	6
4.1. AUTO-BANDWIDTH-ATTRIBUTE TLV . . . . .	6
4.2. BANDWIDTH Object . . . . .	7
4.3. The PCRpt Message . . . . .	8
5. Security Considerations . . . . .	8
6. Manageability Considerations . . . . .	8
6.1. Control of Function and Policy . . . . .	8
6.2. Information and Data Models . . . . .	9
6.3. Liveness Detection and Monitoring . . . . .	9
6.4. Verify Correct Operations . . . . .	9
6.5. Requirements On Other Protocols . . . . .	9
6.6. Impact On Network Operations . . . . .	9
7. IANA Considerations . . . . .	9
7.1. PCEP TLV Type Indicators . . . . .	9
7.2. BANDWIDTH Object . . . . .	9
8. Acknowledgments . . . . .	10
9. References . . . . .	10
9.1. Normative References . . . . .	10
9.2. Informative References . . . . .	10
Appendix A. Contributor Addresses . . . . .	10

## 1. Introduction

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

[STATEFUL-PCE] specifies extensions to PCEP to enable stateful control of MPLS TE LSPs. In this document focus is on Active Stateful PCE where LSPs are configured on the PCC and control over them is delegated to the PCE.

Over time, based on the varying traffic pattern, an LSP established with certain bandwidth may require to adjust the reserved bandwidth over time automatically. Ingress Label Switch Router (LSR) samples the traffic rate at each sample-interval to determine the traffic information as Maximum Average Bandwidth (MaxAvgBw). Further adjustment to the reserved bandwidth should be made at every adjustment-interval automatically.

Enabling Auto-Bandwidth on a LSP results in the LSP automatically adjusting its bandwidth based on the actual traffic flowing through the LSP. A LSP can therefore be setup with some arbitrary (or zero) bandwidth value such that the LSP automatically monitors the traffic flow and adjusts its bandwidth every adjustment-interval period. The bandwidth adjustment uses the make-before-break signaling method so that there is no interruption to traffic flow. This is described in detail in Section 3.1.

[STATEFUL-PCE-APP] describes the usecase for auto-bandwidth adjustment for passive and active stateful PCE. Active stateful PCE can use information such as historical trending data, application-specific information about expected demands or policy information, as well as knowledge of the actual desired flow volumes to make smarter bandwidth adjustment to delegated LSPs.

This document defines extensions needed to support Auto-Bandwidth feature along with mechanism to provide traffic information of the LSPs in a stateful PCE model using PCEP. At the same time this document does not exclude use of any other mechanism employed by stateful PCE to learn real time traffic information etc.

### 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.

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

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

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

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

**TE LSP:** Traffic Engineering Label Switched Path.

Note the additional terms defined in Section 3.1.

## 3. Architectural Overview

### 3.1. Auto-Bandwidth Overview

Auto-Bandwidth feature allows an LSP to automatically and dynamically adjust its reserved bandwidth over time, i.e. without network operator intervention. The bandwidth adjustment uses the make-before-break adaptive signaling method so that there is no interruption to traffic flow.

The new bandwidth reservation is determined by sampling the actual traffic flowing through the LSP. If the traffic flowing through the LSP is lower than the configured or current bandwidth of the LSP, the extra bandwidth is being reserved needlessly and being wasted. Conversely, if the actual traffic flowing through the LSP is higher than the configured or current bandwidth of the LSP, it can potentially cause congestion or packet loss. With Auto-Bandwidth feature, the LSP bandwidth can be set to some arbitrary value (even zero) during initial setup time, and it will be periodically adjusted over time based on the actual bandwidth requirement.

Note the following terms:

**Maximum Average Bandwidth (MaxAvgBw):** The maximum average bandwidth is the unit to measure the current traffic demand between a time interval. This is the maximum value of the averaged traffic pattern in a particular time interval.

**Sample-Interval:** The time interval in which the traffic rate (MaxAvgBw) is collected as a sample.

**Adjustment-Interval:** The time interval in which the bandwidth adjustment should be made based on the MaxAvgBw.

**Minimum Bandwidth:** The minimum bandwidth that should be reserved for the LSP.

**Maximum Bandwidth:** The maximum bandwidth that can be reserved for the LSP.

**Report-Threshold:** This value indicates when the MaxAvgBw must be reported to stateful PCE via PCRpt message. Only if the percentage difference between the current MaxAvgBw and the last MaxAvgBw is greater than or equal to the threshold percentage the LSP bandwidth is reported to PCE.

**Adjust-Threshold:** This value indicates when the bandwidth must be adjusted. Only if the percentage difference between the current MaxAvgBw and the current bandwidth allocation is greater than or equal to the threshold percentage the LSP bandwidth is adjusted to the current bandwidth demand.

### 3.2. Deploying Auto-Bandwidth Feature

The traffic rate is repeatedly sampled at each sample-interval (which can be configured by the user and the default value as 5 minutes). The sampled traffic rates are accumulated over the adjustment-interval period (which can be configured by the user and the default value as 24 hours).

The ingress LSR reports the traffic information to the stateful PCE via the PCRpt message, to avoid multiple reports, the Report-Threshold percentage is used. Only if the percentage difference between the current MaxAvgBw and the last MaxAvgBw is greater than or equal to the threshold percentage the LSP bandwidth is reported to PCE.

Stateful PCE will adjust the bandwidth of the LSP to the highest sampled traffic rate amongst the set of samples taken over the adjustment-interval. Note that the highest sampled traffic rate could be higher or lower than the current LSP bandwidth. Only if the



current MaxAvgBw and the current bandwidth allocation is greater than or equal to the Adjust-Threshold percentage the LSP bandwidth is adjusted to the current bandwidth demand.

Also to avoid multiple LSP re-signaling, sometimes operator set up longer adjustment intervals. However long adjustment-interval can also result in an undesirable effect of masking sudden changes in traffic patterns. To avoid this, the stateful PCE MAY pre-maturely expire the adjustment-interval to accommodate sudden bursts in traffic.

#### 4. Extensions to the PCEP

The extensions are needed to support -

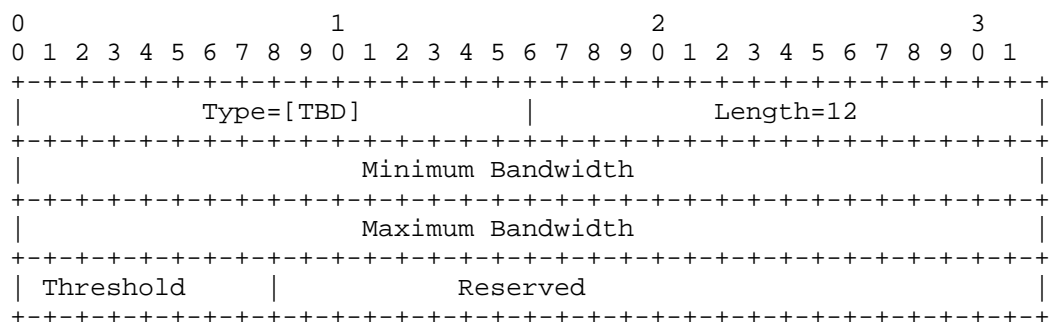
- o identify the LSP that would like to use this feature.
- o specify the auto-bandwidth parameters like bandwidth range.
- o report the live traffic information.

Extensions as specified in this document is one of the way for PCE to learn this information. A stateful PCE MAY learn this information from other means (like NMS, OSS etc).

##### 4.1. AUTO-BANDWIDTH-ATTRIBUTE TLV

The AUTO-BANDWIDTH-ATTRIBUTE TLV can be included as an optional TLV in the LSP object as described in [STATEFUL-PCE]. Whenever the LSP with Auto-Bandwidth feature enabled is delegated, AUTO-BANDWIDTH-ATTRIBUTE TLV MUST be carried in PCRpt message.

The format of the AUTO-BANDWIDTH-ATTRIBUTE TLV is shown in the following figure:



## AUTO-BANDWIDTH-ATTRIBUTE TLV format

The type of the TLV is [TBD] and it has a fixed length of 12 octets.

The value contains the following fields:

Minimum Bandwidth (32 bits): The minimum bandwidth allowed is encoded in IEEE floating point format (see [IEEE.754.1985]), expressed in bytes per second. Refer to Section 3.1.2 of [RFC3471] for a table of commonly used values.

Maximum Bandwidth (32 bits): The maximum bandwidth allowed is encoded in IEEE floating point format (see [IEEE.754.1985]), expressed in bytes per second. Refer to Section 3.1.2 of [RFC3471] for a table of commonly used values.

Threshold (8 bits): The Adjust-Threshold value is encoded in percentage. Only if the percentage difference between the current MaxAvgBw and the current bandwidth allocation is greater than or equal to the threshold percentage the LSP bandwidth is adjusted to the current bandwidth demand.

Reserved (24 bits): These bits MUST be set to zero on transmission and MUST be ignored on receipt.

If the above parameters are not specified by the user, based on the local policy at Ingress (PCC) the default value can be encoded.

If no default value is specified at Ingress, value 'zero' can be encoded for the particular field. The stateful PCE can then apply its own default value based on the local policy.

#### 4.2. BANDWIDTH Object

As per [RFC5440], the BANDWIDTH object is defined with two Object-Type values:

- o Requested Bandwidth: BANDWIDTH Object-Type is 1.
- o Re-optimization Bandwidth: Bandwidth of an existing TE LSP for which a reoptimization is requested. BANDWIDTH Object-Type is 2.

The new BANDWIDTH object type 3 [TBD] is used to specify the MaxAvgBw determined from the existing TE LSP Traffic flow at every sample-interval. The Report-Threshold percentage is used to determine if there is a need to report the current MaxAvgBw.

#### 4.3. The PCRpt Message

When the delegated LSP is enabled with the Auto-Bandwidth adjustment feature, a PCC MAY include the BANDWIDTH object of type 3 [TBD] in the PCRpt message. The definition of the PCRpt message (see [STATEFUL-PCE]) is then extended as follows:

```
<PCRpt Message> ::= <Common Header>
                        <state-report-list>
```

Where:

```
<state-report-list> ::= <state-report>[<state-report-list>]
```

```
<state-report> ::= <LSP>
                    [<path-list>]
```

Where:

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

```
<path> ::= <ERO><attribute-list>
```

Where:

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

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

The BANDWIDTH object of type 3 [TBD] is encoded along with RRO to report the traffic flow information as the MaxAvgBw.

#### 5. Security Considerations

TBD.

#### 6. Manageability Considerations

##### 6.1. Control of Function and Policy

The Auto-Bandwidth feature MUST BE controlled per tunnel at Ingress (PCC), the values for parameters like sample-interval, adjustment-interval, minimum-bandwidth, maximum-bandwidth, report-threshold, adjust-threshold SHOULD BE configurable by user.

TBD.

## 6.2. Information and Data Models

TBD.

## 6.3. Liveness Detection and Monitoring

TBD.

## 6.4. Verify Correct Operations

TBD.

## 6.5. Requirements On Other Protocols

TBD.

## 6.6. Impact On Network Operations

TBD.

## 7. IANA Considerations

### 7.1. PCEP TLV Type Indicators

This document defines the following new PCEP TLVs; IANA is requested to make the following allocations from this registry.

Value	Meaning	Reference
TBD	AUTO-BANDWIDTH-ATTRIBUTE	[This I.D.]

### 7.2. BANDWIDTH Object

This document defines new object type for the BANDWIDTH object; IANA is requested to make the following allocations from this registry.

Object-Class Value	Name	Reference
5	BANDWIDTH	[This I.D.]
	Object-Type	
	3: MaxAvgBw determined from the existing TE LSP Traffic flow.	

## 8. Acknowledgments

We would like to thank Venugopal Reddy, Reeja Paul, Sandeep Boina and Avantika for their useful comments and suggestions.

## 9. References

### 9.1. Normative References

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

### 9.2. Informative References

- [RFC3471] Berger, L., "Generalized Multi-Protocol Label Switching (GMPLS) Signaling Functional Description", RFC 3471, January 2003.
- [RFC5440] Vasseur, JP. and JL. Le Roux, "Path Computation Element (PCE) Communication Protocol (PCEP)", RFC 5440, March 2009.
- [IEEE.754.1985] IEEE, "Standard for Binary Floating-Point Arithmetic".
- [STATEFUL-PCE] Crabbe, E., Medved, J., Minei, I., and R. Varga,, "PCEP Extensions for Stateful PCE (draft-ietf-pce-stateful-pce)", Oct 2012.
- [STATEFUL-PCE-APP] Zhang, F., Zhang, X., Lee, Y., Casellas,, R., and O. Gonzalez de Dios,, "Applicability of Stateful Path Computation Element (PCE) (draft-zhang-pce-stateful-pce-app)".

## Appendix A. Contributor Addresses

He Zekun  
Tencent Holdings Ltd,  
Shenzhen P.R.China

Email: kinghe@tencent.com

Xian Zhang  
Huawei Technologies  
F3-5-B R&D Center, Huawei Base  
Bantian, Longgang District  
Shenzhen 518129 P.R.China

Phone: +86-755-28972913  
Email: zhang.xian@huawei.com

Young Lee  
Huawei  
1700 Alma Drive, Suite 100  
Plano, TX 75075  
US

Phone: +1 972 509 5599 x2240  
Fax: +1 469 229 5397  
EMail: leeyoung@huawei.com

#### Authors' Addresses

Dhruv Dhody  
Huawei Technologies India Pvt Ltd  
Leela Palace  
Bangalore, Karnataka 560008  
INDIA

Email: dhruv.dhody@huawei.com

Udayasree Palle  
Huawei Technologies India Pvt Ltd  
Leela Palace  
Bangalore, Karnataka 560008  
INDIA

Email: udayasree.palle@huawei.com

Path Computation Element Working Group  
Internet-Draft  
Intended status: Informational  
Expires: January 16, 2014

O. Dugeon  
J. Meuric  
Orange Labs  
R. Douville  
Alcatel-Lucent  
R. Casellas  
CTTC  
O. Gonzalez de Dios  
Telefonica Investigacion y Desarrollo  
July 15, 2013

Path Computation Element (PCE) Database Requirements  
draft-dugeon-pce-ted-reqs-02

## 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 and GMPLS LSP tunnels placement. In the PCE architecture, a main assumption has been done concerning the information that the PCE needs to perform its computation. In a first approach, the PCE embeds a Traffic Engineering Database (TED) containing all pertinent and suitable information regarding the network that is in the scope of a PCE. Nevertheless, the TED requirements as well as the TED information have not yet been formalized. In addition, some recent RFC (like the Backward Recursive Path Computation procedure or PCE Hierarchy) or WG draft (like draft-ietf-pce-stateful-pce ...) suffer from a lack of information in the TED, leading to a non optimal result or to some difficulties to deploy them. This memo tries to identify some Database, at large, requirements for the PCE. It is split in two main sections: the identification of the specific information to be stored in the PCE Database and how it may be populated.

## 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/>.

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

This Internet-Draft will expire on January 16, 2014.

#### Copyright Notice

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

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

#### Table of Contents

1. Problem Statement . . . . .	3
1.1. PCE Assumption and Hypothesis . . . . .	3
1.2. Terminology . . . . .	4
2. PCED Requirements . . . . .	5
2.1. Intra-Domain . . . . .	5
2.1.1. MPLS . . . . .	5
2.1.2. GMPLS . . . . .	6
2.2. Inter-Domain . . . . .	6
2.3. TE LSPs . . . . .	6
2.4. Operational Information . . . . .	7
3. PCED model . . . . .	7
3.1. Intra-domain . . . . .	7
3.1.1. MPLS . . . . .	7
3.1.2. GMPLS . . . . .	7
3.2. Inter-domain . . . . .	7
4. PCED Population . . . . .	8
4.1. Intra-domain . . . . .	8
4.1.1. MPLS . . . . .	9
4.1.2. GMPLS . . . . .	9
4.2. Inter-Domain . . . . .	9
4.2.1. Information exchange . . . . .	10



4.3. TE-LSPs . . . . .	11
4.4. Operational information . . . . .	11
5. IANA Considerations . . . . .	11
6. Security Considerations . . . . .	11
6.1. Intra-domain information . . . . .	12
6.2. Inter-domain information . . . . .	12
6.3. Operational information . . . . .	12
7. Acknowledgements . . . . .	12
8. References . . . . .	12
8.1. Normative References . . . . .	12
8.2. Informative References . . . . .	13
Authors' Addresses . . . . .	14

## 1. Problem Statement

Looking to the different RFCs that describe the PCE architecture and in particular RFC 4655 [RFC4655], RFC 5440 [RFC5440], RFC 5441 [RFC5441] and RFC 6805 [RFC6805], the Path Computation Element (PCE) needs to acquire a set of information that is usually store in the Traffic Engineering Database (TED) in order to perform its path computation. Even if intra-domain topology acquisition is well documented and known (e.g. by listening to the IGP-TE protocol that runs inside the network), inter-domain topology information, PCE peer address, neighbor AS, existing MPLS-TE tunnels... that are necessary for the Global Concurrent Optimization, Backward Recursive Path Computation (BRPC) and the Hierarchical PCE are not documented and not completely standardized.

The purpose of this memo is to inventory the required information that should be part of the PCE Database and the different mechanisms that allow an operator to populate it.

### 1.1. PCE Assumption and Hypothesis

In some cases, both the path computation and the Database operations are slightly coupled: border node identification, endpoint localization, TE-LSP learning and domain sequence selection... to name a few in which an IGP-based TED may not be sufficient. It is also important to differentiate several environments with different requirements, especially for the multi-domain problem. The PCE is scoped for any kind of network, from transmission networks (TDM/WDM) with a rather limited number of domains, few interconnections, and few confidentiality issues; transmission networks with a large number of domains; MPLS networks with several administrative domains; and big IP/MPLS networks with a large number of domains with peering agreements. For each of them, a different solution for the multi-domain path computation may apply. A solution may not be scalable for one, but perfectly suitable for another.

Up to now, PCE WG has based its work and standard on the assumption and hypothesis that the TED contains all pertinent information suitable for the PCE to compute an optimal TE-LSP placement, over one or several domains a PCE has visibility on or over a set of PCE-capable domains (e.g. using BRPC procedure). We could identify two major sources of information for the TED:

- o The intra-domain routing protocol like OSPF-TE or IS-IS-TE (including extensions for border links),
- o The inter-domain routing protocol, i.e. BGP for the inter-AS case.

If the first source gives a precise and synchronize view of the controlled network, BGP typically just provides network reachability with only one AS path (unless using recent add path option). Nevertheless, to optimize inter-domain path computation, route diversity and a minimum set of Traffic Engineering information about the foreign domains could be helpful. Despite that it is possible to re-announce TE-LSP in the IGP-TE, the PCE needs also to have a precise knowledge of previous TE-LSP, not only for its stateful version [PCEP Extensions for Stateful PCE] [I-D.ietf-pce-stateful-pce], but also when performing a global concurrent optimization RFC5557 [RFC5557] of the previous TE-LSPs place on a given domain.

Another source of information, mainly static information can be the management plane, e.g. using SNMP, CLI... So, it is necessary to classify the source of information by their frequency of update: static or dynamic, e.g. a domain ID is unlikely to change, while unreserved bandwidth of a link may be continuously changing.

In this document, PCE Database (PCED) is used not only to refer to the standard Traffic Engineering Database information, but is extended to all pertinent information e.g. it also contains previous TE-LSPs establish in the domain and sometimes referred as LSP DB in other documents.

## 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

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

Domain: an Autonomous System

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

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

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

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

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

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).

PCED: Path Computation Element Database

TED: Traffic Engineering Database.

## 2. PCED Requirements

This section made a first inventory of the main requirements of the PCED in term of information that the database should contains.

### 2.1. Intra-Domain

This section describes the Intra-domain information that are suitable for the PCE Database including both MPLS and GMPLS.

#### 2.1.1. MPLS

A PCE is allowed to compute paths in one or several domains. Such PCE MUST be aware of the precise details of the network topology (or topologies) in order to compute optimal TE-LSP placements. The information needed in this case includes:

- o List of Internal Nodes identified by a reachable address: All nodes of the networks that are not border node,

- o List of Internal Links that rely nodes (both internal and border nodes),
- o Traffic Engineering information of the different links i.e. RFC 3630 [RFC3630] and RFC 5305 [RFC5305](with e.g. recent metric extensions proposal OSPF Traffic Engineering (TE) Metric Extensions [I-D.ietf-ospf-te-metric-extensions])
- o Traffic Engineering information with GMPLS extensions of the different links i.e. RFC 4203 [RFC4203] and RFC 5307 [RFC5307],
- o Traffic Engineering information of the nodes.

The information above mentioned is usually exchanged using the IGP-TE protocol (OSPF-TE or IS-IS-TE).

#### 2.1.2. GMPLS

To be provided later

#### 2.2. Inter-Domain

A PCE can also be allowed to take part to inter-domain path computation (e.g in per-domain path computation, BRPC or H-PCE relationship). Some inter-domain information is mandatory when operator intend to use the PCE to compute Inter-AS TE LSP path that cross domain boundary. For that purpose, the PCED SHOULD contains all information that allow the PCE to determine the optimal inter-domain path for the TE-LSP computation, which includes:

- o Border Nodes (BNs) of the foreign domain,
- o Links between BN, i.e. links between BN (n) to BN (n+1), including Traffic Engineering information,
- o Traffic Engineering performance between BN (n) to give performance indication on foreign domain n,
- o PCE (i) peer address associated with the domain number of the foreign domain (i),

RFC 5316 [RFC5316] for IS-IS and RFC 5392 [RFC5392] for OSPF help to provide required PCED information in the case of inter-domain. PCED can also contain information about virtual links and abstract information.

#### 2.3. TE LSPs

For Stateful operation and Global Concurrent Optimization, the PCED should also contain information on TE-LSPs already enforce in the controlled domain. If some TE-LSP tunnels could be re-announce in the IGP-TE, the PCE could not learn from the IGP-TE all details of all TE LSPs: if TE information is known, detail of the ERO is lost as well as initial QoS parameters. The following information will be useful for the PCED to describe the TE-LSP:

- o Explicit Route Object (ERO),
- o End-points objects,
- o Initial and actual Metric objects, including extend metrics such as delay, jitter loss,

Recent PCEP Extensions for Stateful PCE [I-D.ietf-pce-stateful-pce] provide new PCEP message to convey these kind of information. However, this capacity could be used disregarding the behavior (stateless or stateful) of the PCE.

#### 2.4. Operational Information

This part of the TED contains all others information pertinent for the PCE to compute TE LSP path but that are provided through the management system.

### 3. PCED model

This section propose a basic model to store pertinent information regarding the different source of information.

#### 3.1. Intra-domain

##### 3.1.1. MPLS

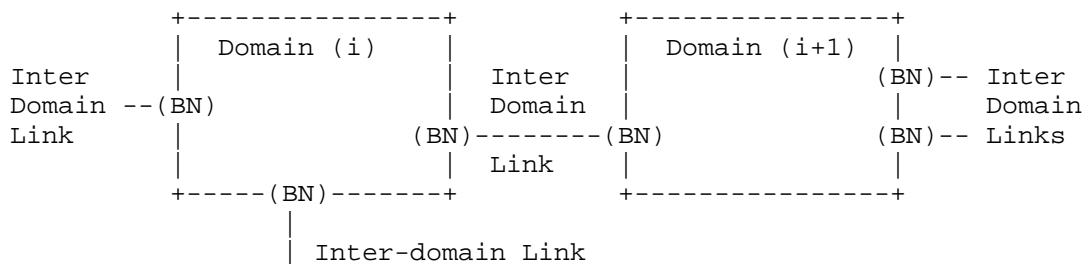
For intra-domain, there is no need to specify a particular model or schema for the PCED. Indeed, the model is directly based on the IGP-TE. Of course there is a difference between IS-IS and OSPF, but TE Link state are more or less similar in term of conveyed information and database description. No particular requirements are necessary as this stage.

##### 3.1.2. GMPLS

To be provided later.

#### 3.2. Inter-domain

Contrary to intra-domain where the PCE known the exact details of the underlying network, it is not possible to achieve a similar detail level for the inter-domain. And not only for scalability reasons, but mostly for confidentiality of the networks. This memo propose a basic schema that allows PCE to know sufficient details about the foreign domain while keeping confidential the internal information. For this purpose, we propose to describe a domain as a "Grey-Box" with inputs and outputs that correspond to the Border Nodes (BNs). Then Grey-Boxes are interconnected through inter-domain links between the BNs. Then, suitable performance indicators are given to cross the Grey-Boxes from an input BN to and output BN. Figure below gives as example of such model.



Example of the representation of 2 domains with the Grey-Box model

With such inter-domain information, a PCE could look into the different inter-domain path (as the sum of inter-domain links and Grey-Box crossing performances) and select the most suitable one regarding the PCReq.

If the inter-domain links between BN that connect the Grey-Boxes description are covered (see section 2.2), it is not the case for the internal links between BNs inside the Grey-Box.

#### 4. PCED Population

This section aims to provide best current practices when mechanisms are well-known and some hints when standard solutions exist to populate the PCE TED, and so give directions to extend them. In particular, we aim at providing input on whether the TED gets the information from the routing protocol and how it gets it, which specific routing protocols are suited, whether it gets it from an NMS, at what frequency the TED is updated... and if it needs extra information.

##### 4.1. Intra-domain

#### 4.1.1. MPLS

As the TED mainly contains the intra-domain topology graph, it is RECOMMENDED to link the PCE with the underlying IGP-TE (OSPF-TE or IS-IS-TE routing protocol). By adding the PCE into the IGP-TE routing intra-domain, it is possible to listen to the routing protocol and then acquired the complete topology graph as well as let the PCE announce itself (see RFC5088 and RFC5089). In addition, the TED will synchronize as fast as the routing protocol converges like any router in the domain. Best current practices are also of interest when a PCE compute path that spawn to several area / region. In that case, the PCE must be aware of the topology details of each area / region and not only the backbone area / region 1 with the summary of stub area / region 2.

In addition, management tools may be used to complement the topology graph provided by the routing protocol.

#### 4.1.2. GMPLS

To be provided later.

#### 4.2. Inter-Domain

If for inter-area aspect of the inter-domain, actual IGP-TE protocol provide the aforementioned information without any particular extension, this is not the case for the inter-as scenario.

First of all, RFC 5316 or RFC 5392 MUST be activated in the IGP-TE (respectively in IS-IS-TE or OSPF-TE) in order to advertise TE information on the inter-domain links. This give the advantage for the PCE to determine what could be feasible, during path computation, on the peering links.

In MPLS, AS path and network reachability are obtained from BGP and routing tables. However, it is not straightforward to collect route diversity or TE information (i.e. bandwidth, transit delay, packet loss ratio, jitter ...) on a foreign domain. Right now, we have identified several methods, which have been tested to fill in the PCED with this kind of information:

- o Use of the management plane;
- o Use of the "North bound distribution of Link-State and TE Information using BGP" [I-D.ietf-idr-ls-distribution] proposal to exchange TE information about the foreign domain;

- o Use of PCNTf message to convey, inside vendor attribute (but in an extended way), TE information of foreign domain between PCE

As well as some potential alternative mechanisms that would need more standardization effort:

- o A hierarchical TE that could help to advertise, at the AS level, TE information on an abstract/aggregate view of the foreign AS topology;
- o A PCEP extension to convey such TE information to the foreign PCE.

#### 4.2.1. Information exchange

The force of PCE is to be aware of the complete topology of the underlying network. With such knowledge, it could place efficiently the tunnel even if it not follows the route computed by the routing protocol. Same principles apply also for the inter-domain. But, in the Internet today, BGP summarize the route and the PCE should not be aware of the route diversity. In particular, it could not choose another AS path as the one selected and announced by BGP. In such case, the PCE will not be sufficiently aware of the route diversity and could not selected the optimal AS path when computing an inter-domain LSP. To avoid this and allows PCE known route diversity to reach a given foreign domain, the inter-domain information must be propagated between all PCEs without aggregation or summarization. In summary, PCEs need to synchronize part of their Database i.e. the inter-domain ones. Disregarding the protocol, two different solutions emerged to exchange inter-domain information:

- o Direct Distribution: Exchange TE information using BGP is part of this case. In this scenario, it is necessary to establish a BGP session between the different domains (whatever the platform used, a dedicated router, a PCE, another server ...). In the hierarchal PCE scenario, operators that provide child PCE, agree to establish a relation with foreign domain that provides the parent PCE. But, in BRPC, or in Hierarchical PCE where almost operators provide a parent PCE, BGP session must be establish between networks that have not necessary direct adjacency. However, operators should not agree to accept relation from other's not directly attached to their network. In addition, this scenario could conduct to establish a full mesh of BGP session between PCE which could lead into some scalability problems.
- o Flooding Distribution: In this case, the inter-domain information are flood between all PCE so that each PCE is aware about all foreign domain capabilities. This meets the requirement but doesn't provide the flexibility of BGP in term of filtering.



Indeed, BGP allows through configuration to decide which information are announced and to whom. As a per session relation, a given operator is not oblige to announce the same capabilities to its foreign domain. With flooding distribution, where everybody redistribute what it has learned without modify it, it is not possible to specialize announcement based on foreign domain.

So, the solution must provide the possibility to filter what is announce per foreign domain without authorized the summarization or aggregation while keeping a distributed relation between domains. In addition, a domain is responsible about the Grey-Box announcement and the advertisement information must not be modified by intermediate PCE.

#### 4.3. TE-LSPs

Up to know, the PCE could learn the tunnel already enforce in the controlled domain through dedicated NMS system. Recent works on state full extensions for PCEP propose to add new messages in order to collect information on TE-LSPs from the PCCs.

#### 4.4. Operational information

Most of the time operational information are provided through the management system of the operator, but some could be automatically discovered. In particular, in intra-domain, PCCs and PCEs can discover automatically reachable PCEs (as well as computation domains) through the deployment of RFC 5088 [RFC5088], for OSPF-controlled networks, and RFC 5089 [RFC5089] for IS-IS controlled networks. However, for the inter-PCE discovery at the inter-AS level, no mechanism has been standardized (unless ASes are owned by the same ISP).

#### 5. IANA Considerations

This document makes no request of IANA for the moment.

Note to RFC Editor: this section may be removed on publication as an RFC.

#### 6. Security Considerations

Acquisition of information for the PCE TED is of course sensible from a security point of view, especially when acquiring information from others AS. This section aims at providing best practices to prevent some security threat when the PCE try to acquire TED information.

### 6.1. Intra-domain information

Same security considerations must be applied to the PCE when it is connected to an IGP-TE protocol as the routing protocol itself. Best practices observed and deployed by operators must also be taken into account when installing some PCEs. Indeed, even when deployed as a standalone server, PCEs must be considered as a typical router from the IGP-TE perspective. As a result, beyond OSPF or IS-IS themselves, the usual security rules must be applied, e.g. login/passwd, authentication/digest... to protect the connectivity.

### 6.2. Inter-domain information

Inter-domain relation and so information exchange are subject to high potential hijack and so need attention from the security point of view. To avoid disclosing or expose confidential information that two operators would exchange to fill in the TEDs of their respective PCEs, the relation SHOULD be protected by standard cryptography mechanism. E.g. using IPsec tunnel is RECOMMENDED to protect the connectivity between PCEs and the TED exchanges.

### 6.3. Operational information

All operational information like PCE peer addresses are generally added manually to the TED and so do not need any particular protection nor subject to security. But, as this basic information is needed to connected the PCEs to their peers, it could potentially be associated to sensitive parameters like login and password. So, standard Best Practices are RECOMMENDED to avoid basic security exposition.

## 7. Acknowledgements

The authors want to thanks PCE's WG members and in particular Daniel King for their inputs of this subject.

## 8. References

### 8.1. Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC4655] Farrel, A., Vasseur, J., and J. Ash, "A Path Computation Element (PCE)-Based Architecture", RFC 4655, August 2006.

- [RFC5440] Vasseur, JP. and JL. Le Roux, "Path Computation Element (PCE) Communication Protocol (PCEP)", RFC 5440, March 2009.
- [RFC5441] Vasseur, JP., Zhang, R., Bitar, N., and JL. Le Roux, "A Backward-Recursive PCE-Based Computation (BRPC) Procedure to Compute Shortest Constrained Inter-Domain Traffic Engineering Label Switched Paths", RFC 5441, April 2009.

## 8.2. Informative References

- [I-D.ietf-idr-ls-distribution]  
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-03 (work in progress), May 2013.
- [I-D.ietf-ospf-te-metric-extensions]  
Giacalone, S., Ward, D., Drake, J., Atlas, A., and S. Previdi, "OSPF Traffic Engineering (TE) Metric Extensions", draft-ietf-ospf-te-metric-extensions-04 (work in progress), June 2013.
- [I-D.ietf-pce-stateful-pce]  
Crabbe, E., Medved, J., Minei, I., and R. Varga, "PCEP Extensions for Stateful PCE", draft-ietf-pce-stateful-pce-05 (work in progress), July 2013.
- [RFC3630] Katz, D., Kompella, K., and D. Yeung, "Traffic Engineering (TE) Extensions to OSPF Version 2", RFC 3630, September 2003.
- [RFC4203] Kompella, K. and Y. Rekhter, "OSPF Extensions in Support of Generalized Multi-Protocol Label Switching (GMPLS)", RFC 4203, October 2005.
- [RFC5088] Le Roux, JL., Vasseur, JP., Ikejiri, Y., and R. Zhang, "OSPF Protocol Extensions for Path Computation Element (PCE) Discovery", RFC 5088, January 2008.
- [RFC5089] Le Roux, JL., Vasseur, JP., Ikejiri, Y., and R. Zhang, "IS-IS Protocol Extensions for Path Computation Element (PCE) Discovery", RFC 5089, January 2008.
- [RFC5305] Li, T. and H. Smit, "IS-IS Extensions for Traffic Engineering", RFC 5305, October 2008.

- [RFC5307] Kompella, K. and Y. Rekhter, "IS-IS Extensions in Support of Generalized Multi-Protocol Label Switching (GMPLS)", RFC 5307, October 2008.
- [RFC5316] Chen, M., Zhang, R., and X. Duan, "ISIS Extensions in Support of Inter-Autonomous System (AS) MPLS and GMPLS Traffic Engineering", RFC 5316, December 2008.
- [RFC5392] Chen, M., Zhang, R., and X. Duan, "OSPF Extensions in Support of Inter-Autonomous System (AS) MPLS and GMPLS Traffic Engineering", RFC 5392, January 2009.
- [RFC5557] Lee, Y., Le Roux, J.L., King, D., and E. Oki, "Path Computation Element Communication Protocol (PCEP) Requirements and Protocol Extensions in Support of Global Concurrent Optimization", RFC 5557, July 2009.
- [RFC6805] King, D. and A. Farrel, "The Application of the Path Computation Element Architecture to the Determination of a Sequence of Domains in MPLS and GMPLS", RFC 6805, November 2012.

#### Authors' Addresses

Olivier Dugeon  
Orange Labs  
2, Avenue Pierre Marzin  
Lannion 22307  
France

Email: [olivier.dugeon@orange.com](mailto:olivier.dugeon@orange.com)

Julien Meuric  
Orange Labs  
2, Avenue Pierre Marzin  
Lannion 22307  
France

Email: [julien.meuric@orange.com](mailto:julien.meuric@orange.com)

Richard Douville  
Alcatel-Lucent  
Route de Villejust  
Nozay 91620  
France

Email: richard.douville@alcatel-lucent.com

Ramon Casellas  
CTTC  
Av. Carl Friedrich FGauss n7  
Castelldefels, Barcelona 08860  
Spain

Email: ramon.casellas@cttc.es

Oscar Gonzalez de Dios  
Telefonica Investigacion y Desarrollo  
C/ Emilio Vargas 6  
Madrid  
Spain

Email: ogondio@tid.es

PCE Working Group  
Internet-Draft  
Intended status: Standards Track  
Expires: March 26, 2017

D. Dhody  
Q. Wu  
Huawei  
V. Manral  
Ionos Network  
Z. Ali  
Cisco Systems  
K. Kumaki  
KDDI Corporation  
September 22, 2016

Extensions to the Path Computation Element Communication Protocol (PCEP)  
to compute service aware Label Switched Path (LSP).  
draft-ietf-pce-pcep-service-aware-13

## Abstract

In certain networks, such as, but not limited to, financial information networks (e.g. stock market data providers), network performance criteria (e.g. latency) are becoming as critical to data path selection as other metrics and constraints. These metrics are associated with the Service Level Agreement (SLA) between customers and service providers. The link bandwidth utilization (the total bandwidth of a link in actual use for the forwarding) is another important factor to consider during path computation.

IGP Traffic Engineering (TE) Metric extensions describe mechanisms with which network performance information is distributed via OSPF and IS-IS respectively. 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 describes the extension to PCEP to carry latency, delay variation, packet loss and link bandwidth utilization as constraints for end to end path computation.

## Status of This Memo

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

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

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any

time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

This Internet-Draft will expire on March 26, 2017.

#### Copyright Notice

Copyright (c) 2016 IETF Trust and the persons identified as the document authors. All rights reserved.

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

#### Table of Contents

1. Introduction . . . . .	3
1.1. Requirements Language . . . . .	4
2. Terminology . . . . .	4
3. PCEP Extensions . . . . .	5
3.1. Extensions to METRIC Object . . . . .	5
3.1.1. Path Delay Metric . . . . .	6
3.1.1.1. Path Delay Metric Value . . . . .	7
3.1.2. Path Delay Variation Metric . . . . .	7
3.1.2.1. Path Delay Variation Metric Value . . . . .	8
3.1.3. Path Loss Metric . . . . .	8
3.1.3.1. Path Loss Metric Value . . . . .	9
3.1.4. Non-Understanding / Non-Support of Service Aware Path Computation . . . . .	9
3.1.5. Mode of Operation . . . . .	10
3.1.5.1. Examples . . . . .	10
3.1.6. Point-to-Multipoint (P2MP) . . . . .	11
3.1.6.1. P2MP Path Delay Metric . . . . .	11
3.1.6.2. P2MP Path Delay Variation Metric . . . . .	11
3.1.6.3. P2MP Path Loss Metric . . . . .	12
3.2. Bandwidth Utilization . . . . .	12
3.2.1. Link Bandwidth Utilization (LBU) . . . . .	12
3.2.2. Link Reserved Bandwidth Utilization (LRBU) . . . . .	12
3.2.3. Bandwidth Utilization (BU) Object . . . . .	13
3.2.3.1. Elements of Procedure . . . . .	14
3.3. Objective Functions . . . . .	15
4. Stateful PCE and PCE Initiated LSPs . . . . .	16

5.	PCEP Message Extension . . . . .	16
5.1.	The PCReq message . . . . .	17
5.2.	The PCRep message . . . . .	17
5.3.	The PCRpt message . . . . .	18
6.	Other Considerations . . . . .	19
6.1.	Inter-domain Path Computation . . . . .	19
6.1.1.	Inter-AS Links . . . . .	19
6.1.2.	Inter-Layer Path Computation . . . . .	19
6.2.	Reoptimizing Paths . . . . .	20
7.	IANA Considerations . . . . .	20
7.1.	METRIC types . . . . .	20
7.2.	New PCEP Object . . . . .	21
7.3.	BU Object . . . . .	21
7.4.	OF Codes . . . . .	22
7.5.	New Error-Values . . . . .	22
8.	Security Considerations . . . . .	22
9.	Manageability Considerations . . . . .	23
9.1.	Control of Function and Policy . . . . .	23
9.2.	Information and Data Models . . . . .	23
9.3.	Liveness Detection and Monitoring . . . . .	23
9.4.	Verify Correct Operations . . . . .	23
9.5.	Requirements On Other Protocols . . . . .	23
9.6.	Impact On Network Operations . . . . .	23
10.	Acknowledgments . . . . .	24
11.	References . . . . .	24
11.1.	Normative References . . . . .	24
11.2.	Informative References . . . . .	25
Appendix A.	PCEP Requirements . . . . .	28
Appendix B.	Contributor Addresses . . . . .	28
Authors' Addresses	. . . . .	29

## 1. Introduction

Real time network performance information is becoming critical in the path computation in some networks. Mechanisms to measure latency, delay variation, and packet loss in an MPLS network are described in [RFC6374]. It is important that latency, delay variation, and packet loss are considered during the path selection process, even before the LSP is set up.

Link bandwidth utilization based on real time traffic along the path is also becoming critical during path computation in some networks. Thus it is important that the link bandwidth utilization is factored in during the path computation.

The Traffic Engineering Database (TED) is populated with network performance information like link latency, delay variation, packet loss, as well as parameters related to bandwidth (residual bandwidth,



available bandwidth and utilized bandwidth) via TE Metric Extensions in OSPF [RFC7471] or IS-IS [RFC7810] or via a management system. [RFC7823] describes how a Path Computation Element (PCE) [RFC4655], can use that information for path selection for explicitly routed LSPs.

A Path Computation Client (PCC) can request a PCE to provide a path meeting end to end network performance criteria. This document extends Path Computation Element Communication Protocol (PCEP) [RFC5440] to handle network performance constraints which include any combination of latency, delay variation, packet loss and bandwidth utilization constraints.

[RFC7471] and [RFC7810] describe various considerations regarding -

- o Announcement thresholds and filters
- o Announcement suppression
- o Announcement periodicity and network stability

The first two provide configurable mechanisms to bound the number of re-advertisements in IGP. The third provides a way to throttle announcements. Section 1.2 of [RFC7823] also describes the oscillation and stability considerations while advertising and considering service aware information.

### 1.1. Requirements Language

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

## 2. Terminology

The following terminology is used in this document.

IGP: Interior Gateway Protocol; Either of the two routing protocols, Open Shortest Path First (OSPF) or Intermediate System to Intermediate System (IS-IS).

IS-IS: Intermediate System to Intermediate System

LBU: Link Bandwidth Utilization (See Section 3.2.1.)

LRBU: Link Reserved Bandwidth Utilization (See Section 3.2.2.)

MPLP: Minimum Packet Loss Path (See Section 3.3.)

MRUP: Maximum Reserved Under-Utilized Path (See Section 3.3.)

MUP: Maximum Under-Utilized Path (See Section 3.3.)

OF: Objective Function; A set of one or more optimization criteria used for the computation of a single path (e.g., path cost minimization) or for the synchronized computation of a set of paths (e.g., aggregate bandwidth consumption minimization, etc). (See [RFC5541].)

OSPF: Open Shortest Path First

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

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

RSVP: Resource Reservation Protocol

TE: Traffic Engineering

TED: Traffic Engineering Database

### 3. PCEP Extensions

This section defines PCEP extensions (see [RFC5440]) for requirements outlined in Appendix A. The proposed solution is used to support network performance and service aware path computation.

#### 3.1. Extensions to METRIC Object

The METRIC object is defined in section 7.8 of [RFC5440], comprising metric-value, metric-type (T field) and a flags field comprising a number of bit-flags (B bit, P bit). This document defines the following types for the METRIC object.

- o T=TBD1: Path Delay metric (Section 3.1.1)
- o T=TBD2: Path Delay Variation metric (Section 3.1.2)
- o T=TBD3: Path Loss metric (Section 3.1.3)
- o T=TBD8: P2MP Path Delay metric (Section 3.1.6.1)
- o T=TBD9: P2MP Path Delay Variation metric (Section 3.1.6.2)

- o T=TBD10: P2MP Path Loss metric (Section 3.1.6.3)

The following terminology is used and expanded along the way.

- o A network comprises of a set of N links  $\{L_i, (i=1\dots N)\}$ .
- o A path P of a point to point (P2P) LSP is a list of K links  $\{L_{pi}, (i=1\dots K)\}$ .

#### 3.1.1. Path Delay Metric

The link delay metric is defined in [RFC7471] and [RFC7810] as "Unidirectional Link Delay". The path delay metric type of the METRIC object in PCEP represents the sum of the link delay metric of all links along a P2P path. Specifically, extending on the above mentioned terminology:

- o A link delay metric of link L is denoted  $D(L)$ .
- o A path delay metric for the P2P path  $P = \text{Sum } \{D(L_{pi}), (i=1\dots K)\}$ .

This is as per the sum of means composition function (section 4.2.5 of [RFC6049]). The section 1.2 of [RFC7823] describes oscillation and stability considerations, and section 2.1 of [RFC7823] describes the calculation of end to end path delay metric. Further section 4.2.9 of [RFC6049] states when this composition function may fail.

Metric Type T=TBD1: Path Delay metric

A PCC MAY use the path delay metric in a PCReq message to request a path meeting the end to end latency requirement. In this case, the B bit MUST be set to suggest a bound (a maximum) for the path delay metric that must not be exceeded for the PCC to consider the computed path as acceptable. The path delay metric must be less than or equal to the value specified in the metric-value field.

A PCC can also use this metric to ask PCE to optimize the path delay during path computation. In this case, the B bit MUST be cleared.

A PCE MAY use the path delay metric in a PCRep message along with a NO-PATH object in the case where the PCE cannot compute a path meeting this constraint. A PCE can also use this metric to send the computed path delay metric to the PCC.

#### 3.1.1.1. Path Delay Metric Value

[RFC7471] and [RFC7810] define the "Unidirectional Link Delay Sub-TLV" to advertise the link delay in microseconds in a 24-bit field. [RFC5440] defines the METRIC object with a 32-bit metric value encoded in IEEE floating point format (see [IEEE.754.1985]). Consequently, the encoding for the path delay metric value is quantified in units of microseconds and encoded in IEEE floating point format. The conversion from 24 bit integer to 32 bit IEEE floating point could introduce some loss of precision.

#### 3.1.1.2. Path Delay Variation Metric

The link delay variation metric is defined in [RFC7471] and [RFC7810] as "Unidirectional Delay Variation". The path delay variation metric type of the METRIC object in PCEP encodes the sum of the link delay variation metric of all links along the path. Specifically, extending on the above mentioned terminology:

- o A delay variation of link L is denoted DV(L) (average delay variation for link L).
- o A path delay variation metric for the P2P path P =  $\text{Sum} \{DV(L_{pi}), (i=1...K)\}$ .

The section 1.2 of [RFC7823] describes oscillation and stability considerations, and section 2.1 of [RFC7823] describes the calculation of end to end path delay variation metric. Further section 4.2.9 of [RFC6049] states when this composition function may fail.

Note that the IGP advertisement for link attributes includes the average delay variation over a period of time. An implementation, therefore, MAY use the sum of the average delay variation of links along a path to derive the delay variation of the path. An end-to-end bound on delay variation is typically used as constraint in the path computation. An implementation MAY also use some enhanced composition function for computing the delay variation of a path with better accuracy.

Metric Type T=TBD2: Path Delay Variation metric

A PCC MAY use the path delay variation metric in a PCReq message to request a path meeting the path delay variation requirement. In this case, the B bit MUST be set to suggest a bound (a maximum) for the path delay variation metric that must not be exceeded for the PCC to consider the computed path as acceptable. The path delay variation

must be less than or equal to the value specified in the metric-value field.

A PCC can also use this metric to ask the PCE to optimize the path delay variation during path computation. In this case, the B flag MUST be cleared.

A PCE MAY use the path delay variation metric in PCRep message along with a NO-PATH object in the case where the PCE cannot compute a path meeting this constraint. A PCE can also use this metric to send the computed end to end path delay variation metric to the PCC.

#### 3.1.2.1. Path Delay Variation Metric Value

[RFC7471] and [RFC7810] define "Unidirectional Delay Variation Sub-TLV" to advertise the link delay variation in microseconds in a 24-bit field. [RFC5440] defines the METRIC object with a 32-bit metric value encoded in IEEE floating point format (see [IEEE.754.1985]). Consequently, the encoding for the path delay variation metric value is quantified in units of microseconds and encoded in IEEE floating point format. The conversion from 24 bit integer to 32 bit IEEE floating point could introduce some loss of precision.

#### 3.1.3. Path Loss Metric

[RFC7471] and [RFC7810] define "Unidirectional Link Loss". The path loss (as a packet percentage) metric type of the METRIC object in PCEP encodes a function of the unidirectional loss metrics of all links along a P2P path. The end to end packet loss for the path is represented by this metric. Specifically, extending on the above mentioned terminology:

- o The percentage link loss of link L is denoted  $PL(L)$ .
- o The fractional link loss of link L is denoted  $FL(L) = PL(L)/100$ .
- o The percentage path loss metric for the P2P path  $P = (1 - ((1-FL(Lp1)) * (1-FL(Lp2)) * .. * (1-FL(LpK)))) * 100$  for a path P with links  $Lp1$  to  $LpK$ .

This is as per the composition function described in section 5.1.5 of [RFC6049].

Metric Type T=TBD3: Path Loss metric

A PCC MAY use the path loss metric in a PCReq message to request a path meeting the end to end packet loss requirement. In this case,

the B bit MUST be set to suggest a bound (a maximum) for the path loss metric that must not be exceeded for the PCC to consider the computed path as acceptable. The path loss metric must be less than or equal to the value specified in the metric-value field.

A PCC can also use this metric to ask the PCE to optimize the path loss during path computation. In this case, the B flag MUST be cleared.

A PCE MAY use the path loss metric in a PCRep message along with a NO-PATH object in the case where the PCE cannot compute a path meeting this constraint. A PCE can also use this metric to send the computed end to end path loss metric to the PCC.

#### 3.1.3.1. Path Loss Metric Value

[RFC7471] and [RFC7810] define "Unidirectional Link Loss Sub-TLV" to advertise the link loss in percentage in a 24-bit field. [RFC5440] defines the METRIC object with 32-bit metric value encoded in IEEE floating point format (see [IEEE.754.1985]). Consequently, the encoding for the path loss metric value is quantified as a percentage and encoded in IEEE floating point format.

#### 3.1.4. Non-Understanding / Non-Support of Service Aware Path Computation

If a PCE receives a PCReq message containing a METRIC object with a type defined in this document, and the PCE does not understand or support that metric type, and the P bit is clear in the METRIC object header then the PCE SHOULD simply ignore the METRIC object as per the processing specified in [RFC5440].

If the PCE does not understand the new METRIC type, and the P bit is set in the METRIC object header, then the PCE MUST send a PCErr message containing a PCEP-ERROR Object with Error-Type = 4 (Not supported object) and Error-value = 4 (Unsupported parameter) [RFC5440][RFC5441].

If the PCE understands but does not support the new METRIC type, and the P bit is set in the METRIC object header, then the PCE MUST send a PCErr message containing a PCEP-ERROR Object with Error-Type = 4 (Not supported object) with Error-value = TBD11 (Unsupported network performance constraint). The path computation request MUST then be cancelled.

If the PCE understands the new METRIC type, but the local policy has been configured on the PCE to not allow network performance constraint, and the P bit is set in the METRIC object header, then

the PCE MUST send a PCErr message containing a PCEP-ERROR Object with Error-Type = 5 (Policy violation) with Error-value = TBD12 (Not allowed network performance constraint). The path computation request MUST then be cancelled.

### 3.1.5. Mode of Operation

As explained in [RFC5440], the METRIC object is optional and can be used for several purposes. In a PCReq message, a PCC MAY insert one or more METRIC objects:

- o To indicate the metric that MUST be optimized by the path computation algorithm (path delay, path delay variation or path loss).
- o To indicate a bound on the METRIC (path delay, path delay variation or path loss) that MUST NOT be exceeded for the path to be considered as acceptable by the PCC.

In a PCRep message, the PCE MAY insert the METRIC object with an Explicit Route Object (ERO) so as to provide the METRIC (path delay, path delay variation or path loss) for the computed path. The PCE MAY also insert the METRIC object with a NO-PATH object to indicate that the metric constraint could not be satisfied.

The path computation algorithmic aspects used by the PCE to optimize a path with respect to a specific metric are outside the scope of this document.

All the rules of processing the METRIC object as explained in [RFC5440] are applicable to the new metric types as well.

#### 3.1.5.1. Examples

If a PCC sends a path computation request to a PCE where the metric to optimize is the path delay and the path loss must not exceed the value of M, then two METRIC objects are inserted in the PCReq message:

- o First METRIC object with B=0, T=TBD1, C=1, metric-value=0x0000
- o Second METRIC object with B=1, T=TBD3, metric-value=M

As per [RFC5440], if a path satisfying the set of constraints can be found by the PCE and there is no policy that prevents the return of the computed metric, then the PCE inserts one METRIC object with B=0, T=TBD1, metric-value= computed path delay. Additionally, the PCE MAY

insert a second METRIC object with B=1, T=TBD3, metric-value=computed path loss.

### 3.1.6. Point-to-Multipoint (P2MP)

This section defines the following types for the METRIC object to be used for the P2MP TE LSPs.

#### 3.1.6.1. P2MP Path Delay Metric

The P2MP path delay metric type of the METRIC object in PCEP encodes the path delay metric for the destination that observes the worst delay metric among all destinations of the P2MP tree. Specifically, extending on the above mentioned terminology:

- o A P2MP tree T comprises a set of M destinations {Dest\_j, (j=1...M)}.
- o The P2P path delay metric of the path to destination Dest\_j is denoted by PDM(Dest\_j).
- o The P2MP path delay metric for the P2MP tree T = Maximum {PDM(Dest\_j), (j=1...M)}.

The value for the P2MP path delay metric type (T) = TBD8.

#### 3.1.6.2. P2MP Path Delay Variation Metric

The P2MP path delay variation metric type of the METRIC object in PCEP encodes the path delay variation metric for the destination that observes the worst delay variation metric among all destinations of the P2MP tree. Specifically, extending on the above mentioned terminology:

- o A P2MP tree T comprises a set of M destinations {Dest\_j, (j=1...M)}.
- o The P2P path delay variation metric of the path to the destination Dest\_j is denoted by PDVM(Dest\_j).
- o The P2MP path delay variation metric for the P2MP tree T = Maximum {PDVM(Dest\_j), (j=1...M)}.

The value for the P2MP path delay variation metric type (T) = TBD9.



### 3.1.6.3. P2MP Path Loss Metric

The P2MP path loss metric type of the METRIC object in PCEP encodes the path packet loss metric for the destination that observes the worst packet loss metric among all destinations of the P2MP tree. Specifically, extending on the above mentioned terminology:

- o A P2MP tree T comprises of a set of M destinations {Dest\_j, (j=1...M)}.
- o The P2P path loss metric of the path to destination Dest\_j is denoted by PLM(Dest\_j).
- o The P2MP path loss metric for the P2MP tree T = Maximum {PLM(Dest\_j), (j=1...M)}.

The value for the P2MP path loss metric type (T) = TBD10.

## 3.2. Bandwidth Utilization

### 3.2.1. Link Bandwidth Utilization (LBU)

The Link Bandwidth Utilization (LBU) on a link, forwarding adjacency, or bundled link is populated in the TED ("Unidirectional Utilized Bandwidth" in [RFC7471] and [RFC7810]). For a link or forwarding adjacency, the bandwidth utilization represents the actual utilization of the link (i.e., as measured in the router). For a bundled link, the bandwidth utilization is defined to be the sum of the component link bandwidth utilization. This includes traffic for both RSVP-TE and non-RSVP-TE label switched path packets.

The LBU in percentage is described as the (utilized bandwidth / maximum bandwidth) \* 100.

Where "maximum bandwidth" is defined in [RFC3630] and [RFC5305] and "utilized bandwidth" in [RFC7471] and [RFC7810].

### 3.2.2. Link Reserved Bandwidth Utilization (LRBU)

The Link Reserved Bandwidth Utilization (LRBU) on a link, forwarding adjacency, or bundled link can be calculated from the TED. The utilized bandwidth includes traffic for both RSVP-TE and non-RSVP-TE LSPs, the reserved bandwidth utilization considers only the RSVP-TE LSPs.

The reserved bandwidth utilization can be calculated by using the residual bandwidth, the available bandwidth and utilized bandwidth described in [RFC7471] and [RFC7810]. The actual bandwidth by non-

RSVP-TE traffic can be calculated by subtracting the available bandwidth from the residual bandwidth ([RFC7471] and [RFC7810]). Which is further deducted from utilized bandwidth to get the reserved bandwidth utilization. Thus,

reserved bandwidth utilization = utilized bandwidth - (residual bandwidth - available bandwidth)

The LRBW in percentage is described as the (reserved bandwidth utilization / maximum reservable bandwidth) \* 100.

Where the "maximum reservable bandwidth" is defined in [RFC3630] and [RFC5305]. The "utilized bandwidth", "residual bandwidth", and "available bandwidth" are defined in [RFC7471] and [RFC7810].

### 3.2.3. Bandwidth Utilization (BU) Object

The BU object is used to indicate the upper limit of the acceptable link bandwidth utilization percentage.

The BU object MAY be carried within the PCReq message and PCRep messages.

BU Object-Class is TBD4.

BU Object-Type is 1.

The format of the BU object body is as follows:

0																								1																								2																								3																							
0	1	2	3	4	5	6	7	8	9	0	1	2	3	4	5	6	7	8	9	0	1	2	3	4	5	6	7	8	9	0	1	2	3	4	5	6	7	8	9																																																								
Reserved																								Type																								Bandwidth Utilization																																															

#### BU Object Body Format

Reserved (24 bits): This field MUST be set to zero on transmission and MUST be ignored on receipt.

Type (8 bits): Represents the bandwidth utilization type. Two values are currently defined.

\* Type 1 is Link Bandwidth Utilization (LBU)

- \* Type 2 is Link Reserved Bandwidth Utilization (LRBU)

Bandwidth Utilization (32 bits): Represents the bandwidth utilization quantified as a percentage (as described in Section 3.2.1 and Section 3.2.2) and encoded in IEEE floating point format (see [IEEE.754.1985]).

The BU object body has a fixed length of 8 bytes.

#### 3.2.3.1. Elements of Procedure

A PCC that wants the PCE to factor in the bandwidth utilization during path computation includes a BU object in the PCReq message. A PCE that supports this object MUST ensure that no link on the computed path has the LBU or LRBU percentage exceeding the given value.

A PCReq or PCRep message MAY contain multiple BU objects so long as each is for a different bandwidth utilization type. If a message contains more than one BU object with the same bandwidth utilization type, the first MUST be processed by the receiver and subsequent instances MUST be ignored.

If the BU object is unknown/unsupported, the PCE is expected to follow procedures defined in [RFC5440]. That is, if the P bit is set, the PCE sends a PCErr message with error type 3 or 4 (Unknown / Not supported object) and error value 1 or 2 (unknown / unsupported object class / object type), and the related path computation request will be discarded. If the P bit is cleared, the PCE is free to ignore the object.

If the PCE understands but does not support path computation requests using the BU object, and the P bit is set in the BU object header, then the PCE MUST send a PCErr message with a PCEP-ERROR Object Error-Type = 4 (Not supported object) with Error-value = TBD11 (Unsupported network performance constraint) and the related path computation request MUST be discarded.

If the PCE understands the BU object but the local policy has been configured on the PCE to not allow network performance constraint, and the P bit is set in the BU object header, then the PCE MUST send a PCErr message with a PCEP-ERROR Object Error-Type = 5 (Policy Violation) with Error-value = TBD12 (Not allowed network performance constraint). The path computation request MUST then be cancelled.

If path computation is unsuccessful, then a PCE MAY insert a BU object (along with a NO-PATH object) into a PCRep message to indicate the constraints that could not be satisfied.

Usage of the BU object for P2MP LSPs is outside the scope of this document.

### 3.3. Objective Functions

[RFC5541] defines a mechanism to specify an objective function that is used by a PCE when it computes a path. The new metric types for path delay and path delay variation can continue to use the existing objective function - Minimum Cost Path (MCP) [RFC5541]. For path loss, the following new OF is defined.

- o A network comprises a set of  $N$  links  $\{L_i, (i=1\dots N)\}$ .
- o A path  $P$  is a list of  $K$  links  $\{L_{p_i}, (i=1\dots K)\}$ .
- o The percentage link loss of link  $L$  is denoted  $PL(L)$ .
- o The fractional link loss of link  $L$  is denoted  $FL(L) = PL(L) / 100$ .
- o The percentage path loss of a path  $P$  is denoted  $PL(P)$ , where  $PL(P) = (1 - ((1-FL(L_{p1})) * (1-FL(L_{p2})) * \dots * (1-FL(L_{pK})))) * 100$ .

Objective Function Code: TBD5

Name: Minimum Packet Loss Path (MPLP)

Description: Find a path  $P$  such that  $PL(P)$  is minimized.

Two additional objective functions -- namely, MUP (the Maximum Under-Utilized Path) and MRUP (the Maximum Reserved Under-Utilized Path) are needed to optimize bandwidth utilization. These two new objective function codes are defined below.

These objective functions are formulated using the following additional terminology:

- o The bandwidth utilization on link  $L$  is denoted  $u(L)$ .
- o The reserved bandwidth utilization on link  $L$  is denoted  $ru(L)$ .
- o The maximum bandwidth on link  $L$  is denoted  $M(L)$ .
- o The maximum reservable bandwidth on link  $L$  is denoted  $R(L)$ .

The description of the two new objective functions is as follows.

Objective Function Code: TBD6  
Name: Maximum Under-Utilized Path (MUP)  
Description: Find a path P such that  $(\text{Min } \{(M(Lpi) - u(Lpi)) / M(Lpi), i=1 \dots K\})$  is maximized.

Objective Function Code: TBD7  
Name: Maximum Reserved Under-Utilized Path (MRUP)  
Description: Find a path P such that  $(\text{Min } \{(R(Lpi) - ru(Lpi)) / R(Lpi), i=1 \dots K\})$  is maximized.

These new objective functions are used to optimize paths based on the bandwidth utilization as the optimization criteria.

If the objective functions defined in this document are unknown/unsupported by a PCE, then the procedure as defined in section 3.1.1 of [RFC5541] is followed.

#### 4. Stateful PCE and PCE Initiated LSPs

[STATEFUL-PCE] specifies a set of extensions to PCEP to enable stateful control of MPLS-TE and GMPLS LSPs via PCEP and maintaining of these LSPs at the stateful PCE. It further distinguishes between an active and a passive stateful PCE. A passive stateful PCE uses LSP state information learned from PCCs to optimize path computations but does not actively update LSP state. In contrast, an active stateful PCE utilizes the LSP delegation mechanism to update LSP parameters in those PCCs that delegated control over their LSPs to the PCE. [PCE-INITIATED] describes the setup, maintenance and teardown of PCE-initiated LSPs under the stateful PCE model. The document defines the PCInitiate message that is used by a PCE to request a PCC to set up a new LSP.

The new metric type and objective functions defined in this document can also be used with the stateful PCE extensions. The format of PCEP messages described in [STATEFUL-PCE] and [PCE-INITIATED] uses <attribute-list> (which is extended in Section 5.2) for the purpose of including the service aware parameters.

The stateful PCE implementation MAY use the extension of PCReq and PCRep messages as defined in Section 5.1 and Section 5.2 to enable the use of service aware parameters during passive stateful operations.

#### 5. PCEP Message Extension

Message formats in this document are expressed using Reduced BNF as used in [RFC5440] and defined in [RFC5511].

### 5.1. The PCReq message

The extensions to PCReq message are -

- o new metric types using existing METRIC object
- o a new optional BU object
- o new objective functions using existing OF object ([RFC5541])

The format of the PCReq message (with [RFC5541] and [STATEFUL-PCE] as a base) is updated as follows:

```

<PCReq Message> ::= <Common Header>
                    [<svec-list>]
                    <request-list>
where:
    <svec-list> ::= <SVEC>
                  [<OF>]
                  [<metric-list>]
                  [<svec-list>]

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

    <request> ::= <RP>
                 <END-POINTS>
                 [<LSP>]
                 [<LSPA>]
                 [<BANDWIDTH>]
                 [<bu-list>]
                 [<metric-list>]
                 [<OF>]
                 [<RRO>[<BANDWIDTH>]]
                 [<IRO>]
                 [<LOAD-BALANCING>]

    and where:
        <bu-list> ::= <BU> [<bu-list>]
        <metric-list> ::= <METRIC> [<metric-list>]

```

### 5.2. The PCRep message

The extensions to PCRep message are -

- o new metric types using existing METRIC object
- o a new optional BU object (during unsuccessful path computation, to indicate the bandwidth utilization as a reason for failure)

- o new objective functions using existing OF object ([RFC5541])

The format of the PCRep message (with [RFC5541] and [STATEFUL-PCE] as a base) is updated as follows:

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

where:

```
<svec-list> ::= <SVEC>
                [<OF>]
                [<metric-list>]
                [<svec-list>]

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

<response> ::= <RP>
                [<LSP>]
                [<NO-PATH>]
                [<attribute-list>]
                [<path-list>]

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

<path> ::= <ERO>
           <attribute-list>
```

and where:

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

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

### 5.3. The PCRpt message

A Path Computation LSP State Report message (also referred to as PCRpt message) is a PCEP message sent by a PCC to a PCE to report the current state or delegate control of an LSP. The BU object in a PCRpt message specifies the upper limit set at the PCC at the time of LSP delegation to an active stateful PCE.

The format of the PCRpt message is described in [STATEFUL-PCE] which uses the <attribute-list> as defined in [RFC5440] and extended by PCEP extensions.

The PCRpt message can use the updated <attribute-list> (as extended in Section 5.2) for the purpose of including the BU object.

## 6. Other Considerations

### 6.1. Inter-domain Path Computation

[RFC5441] describes the Backward Recursive PCE-Based Computation (BRPC) procedure to compute end to end optimized inter-domain path by cooperating PCEs. The new metric types defined in this document can be applied to end to end path computation, in a similar manner to the existing IGP or TE metrics. The new BU object defined in this document can be applied to end to end path computation, in a similar manner to a METRIC object with its B bit set to 1.

All domains should have the same understanding of the METRIC (path delay variation etc.) and the BU object for end-to-end inter-domain path computation to make sense. Otherwise, some form of metric normalization as described in [RFC5441] MUST be applied.

#### 6.1.1. Inter-AS Links

The IGP in each neighbour domain can advertise its inter-domain TE link capabilities. This has been described in [RFC5316] (IS-IS) and [RFC5392] (OSPF). The network performance link properties are described in [RFC7471] and [RFC7810]. The same properties must be advertised using the mechanism described in [RFC5392] (OSPF) and [RFC5316] (IS-IS).

#### 6.1.2. Inter-Layer Path Computation

[RFC5623] provides a framework for PCE-Based inter-layer MPLS and GMPLS Traffic Engineering. Lower-layer LSPs that are advertised as TE links into the higher-layer network form a Virtual Network Topology (VNT). The advertisement into the higher-layer network should include network performance link properties based on the end to end metric of the lower-layer LSP. Note that the new metrics defined in this document are applied to end to end path computation, even though the path may cross multiple layers.



## 6.2. Reoptimizing Paths

[RFC6374] defines the measurement of loss, delay, and related metrics over LSPs. A PCC can utilize these measurement techniques. In case it detects a degradation of network performance parameters relative to the value of the constraint it gave when the path was set up, or relative to an implementation-specific threshold, it MAY ask the PCE to reoptimize the path by sending a PCReq with the R bit set in the RP object, as per [RFC5440].

A PCC may also detect the degradation of an LSP without making any direct measurements, by monitoring the TED (as populated by the IGP) for changes in the network performance parameters of the links that carry its LSPs. The PCC can issue a reoptimization request for any impacted LSPs. For example, a PCC can monitor the link bandwidth utilization along the path by monitoring changes in the bandwidth utilization parameters of one or more links on the path in the TED. If the bandwidth utilization percentage of any of the links in the path changes to a value less than that required when the path was set up, or otherwise less than an implementation-specific threshold, then the PCC can issue an reoptimization request to a PCE.

A stateful PCE can also determine which LSPs should be re-optimized based on network events or triggers from external monitoring systems. For example, when a particular link deteriorates and its loss increases, this can trigger the stateful PCE to automatically determine which LSP are impacted and should be reoptimized.

## 7. IANA Considerations

### 7.1. METRIC types

IANA maintains the "Path Computation Element Protocol (PCEP) Numbers" at <<http://www.iana.org/assignments/pcep>>. Within this registry IANA maintains one sub-registry for "METRIC object T field". Six new metric types are defined in this document for the METRIC object (specified in [RFC5440]).

IANA is requested to make the following allocations:

Value	Description	Reference
TBD1	Path Delay metric	[This I.D.]
TBD2	Path Delay Variation metric	[This I.D.]
TBD3	Path Loss metric	[This I.D.]
TBD8	P2MP Path Delay metric	[This I.D.]
TBD9	P2MP Path Delay variation metric	[This I.D.]
TBD10	P2MP Path Loss metric	[This I.D.]

## 7.2. New PCEP Object

IANA maintains object class in the registry of PCEP Objects at the sub-registry "PCEP Objects". One new allocation is requested as follows.

Object Class	Object Type	Name	Reference
TBD4	1	BU	[This I.D.]

## 7.3. BU Object

This document requests that a new sub-registry, named "BU Object Type Field", is created within the "Path Computation Element Protocol (PCEP) Numbers" registry to manage the Type field of the BU object. New values are to be assigned by Standards Action [RFC5226]. Each value should be tracked with the following qualities:

- o Type
- o Name
- o Defining RFC

The following values are defined in this document:

Type	Name	Reference
1	LBU (Link Bandwidth Utilization)	[This I.D.]
2	LRBU (Link Residual Bandwidth Utilization)	[This I.D.]

#### 7.4. OF Codes

IANA maintains registry of Objective Function (described in [RFC5541]) at the sub-registry "Objective Function". Three new Objective Functions have been defined in this document.

IANA is requested to make the following allocations:

Code Point	Name	Reference
TBD5	Minimum Packet Loss Path (MPLP)	[This I.D.]
TBD6	Maximum Under-Utilized Path (MUP)	[This I.D.]
TBD7	Maximum Reserved Under-Utilized Path (MRUP)	[This I.D.]

#### 7.5. New Error-Values

IANA maintains a registry of Error-Types and Error-values for use in PCEP messages. This is maintained as the "PCEP-ERROR Object Error Types and Values" sub-registry of the "Path Computation Element Protocol (PCEP) Numbers" registry.

IANA is requested to make the following allocations -

Two new Error-values are defined for the Error-Type "Not supported object" (type 4) and "Policy violation" (type 5).

Error-Type	Meaning and error values	Reference
4	Not supported object	
	Error-value=TBD11 Unsupported network performance constraint	[This I.D.]
5	Policy violation	
	Error-value=TBD12 Not allowed network performance constraint	[This I.D.]

#### 8. Security Considerations

This document defines new METRIC types, a new BU object, and new OF codes which does not add any new security concerns beyond those discussed in [RFC5440] and [RFC5541] in itself. Some deployments may find the service aware information like delay and packet loss to be

extra sensitive and could be used to influence path computation and 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 [PCEPS]. The Transport Layer Security (TLS) based procedure in [PCEPS] is considered as a security enhancement and thus much better suited for the sensitive service aware information.

## 9. Manageability Considerations

### 9.1. Control of Function and Policy

The only configurable item is the support of the new constraints on a PCE which MAY be controlled by a policy module on individual basis. If the new constraint is not supported/allowed on a PCE, it MUST send a PCErr message accordingly.

### 9.2. Information and Data Models

[RFC7420] describes the PCEP MIB. There are no new MIB Objects for this document.

### 9.3. 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].

### 9.4. 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].

### 9.5. Requirements On Other Protocols

The PCE requires the TED to be populated with network performance information like link latency, delay variation, packet loss, and utilized bandwidth. This mechanism is described in [RFC7471] and [RFC7810].

### 9.6. 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].

## 10. Acknowledgments

We would like to thank Alia Atlas, John E Drake, David Ward, Young Lee, Venugopal Reddy, Reeja Paul, Sandeep Kumar Boina, Suresh Babu, Quintin Zhao, Chen Huaimo, Avantika, and Adrian Farrel for their useful comments and suggestions.

Also the authors gratefully acknowledge reviews and feedback provided by Qin Wu, Alfred Morton and Paul Aitken during performance directorate review.

Thanks to Jonathan Hardwick for shepherding this document and providing valuable comments. His help in fixing the editorial and grammatical issues is also appreciated.

Thanks to Christian Hopps for the routing directorate review.

Thanks to Jouni Korhonen and Alfred Morton for the operational directorate review.

Thanks to Christian Huitema for the security directorate review.

Thanks to Deborah Brungard for being the responsible AD.

Thanks to Ben Campbell, Joel Jaeggli, Stephen Farrell, Kathleen Moriarty, Spencer Dawkins, Mirja Kuehlewind, Jari Arkko and Alia Atlas for the IESG reviews.

## 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>>.
- [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>>.
- [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>>.

- [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>>.
- [RFC5511] Farrel, A., "Routing Backus-Naur Form (RBNF): A Syntax Used to Form Encoding Rules in Various Routing Protocol Specifications", RFC 5511, DOI 10.17487/RFC5511, April 2009, <<http://www.rfc-editor.org/info/rfc5511>>.
- [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>>.
- [RFC7471] Giacalone, S., Ward, D., Drake, J., Atlas, A., and S. Previdi, "OSPF Traffic Engineering (TE) Metric Extensions", RFC 7471, DOI 10.17487/RFC7471, March 2015, <<http://www.rfc-editor.org/info/rfc7471>>.
- [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>>.
- [STATEFUL-PCE]  
Crabbe, E., Minei, I., Medved, J., and R. Varga, "PCEP Extensions for Stateful PCE", draft-ietf-pce-stateful-pce-16 (work in progress), September 2016.

## 11.2. Informative References

- [RFC4655] Farrel, A., Vasseur, J., and J. Ash, "A Path Computation Element (PCE)-Based Architecture", RFC 4655, DOI 10.17487/RFC4655, August 2006, <<http://www.rfc-editor.org/info/rfc4655>>.
- [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>>.
- [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>>.
- [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>>.
- [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>>.
- [RFC6049] Morton, A. and E. Stephan, "Spatial Composition of Metrics", RFC 6049, DOI 10.17487/RFC6049, January 2011, <<http://www.rfc-editor.org/info/rfc6049>>.
- [RFC6374] Frost, D. and S. Bryant, "Packet Loss and Delay Measurement for MPLS Networks", RFC 6374, DOI 10.17487/RFC6374, September 2011, <<http://www.rfc-editor.org/info/rfc6374>>.
- [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>>.
- [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, DOI 10.17487/RFC7823, May 2016, <<http://www.rfc-editor.org/info/rfc7823>>.
- [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-07 (work in progress), July 2016.

- [PCEPS] Lopez, D., Dios, O., Wu, W., and D. Dhody, "Secure Transport for PCEP", draft-ietf-pce-pceps-10 (work in progress), July 2016.
- [IEEE.754.1985]  
IEEE, "Standard for Binary Floating-Point Arithmetic",  
IEEE 754, August 1985.



## Appendix A. PCEP Requirements

End-to-end service optimization based on latency, delay variation, packet loss, and link bandwidth utilization are key requirements for service providers. The following associated key requirements are identified for PCEP:

1. A PCE supporting this draft MUST have the capability to compute end-to-end (E2E) paths with latency, delay variation, packet loss, and bandwidth utilization constraints. It MUST also support the combination of network performance constraints (latency, delay variation, loss...) with existing constraints (cost, hop-limit...).
2. A PCC MUST be able to specify any network performance constraint in a Path Computation Request (PCReq) message to be applied during the path computation.
3. A PCC MUST be able to request that a PCE optimizes a path using any network performance criteria.
4. A PCE that supports this specification is not required to provide service aware path computation to any PCC at any time. Therefore, it MUST be possible for a PCE to reject a PCReq message with a reason code that indicates service-aware path computation is not supported. Furthermore, a PCE that does not support this specification will either ignore or reject such requests using pre-existing mechanisms, therefore the requests MUST be identifiable to legacy PCEs and rejections by legacy PCEs MUST be acceptable within this specification.
5. A PCE SHOULD be able to return end to end network performance information of the computed path in a Path Computation Reply (PCRep) message.
6. A PCE SHOULD be able to compute multi-domain (e.g., Inter-AS, Inter-Area or Multi-Layer) service aware paths.

Such constraints are only meaningful if used consistently: for instance, if the delay of a computed path segment is exchanged between two PCEs residing in different domains, a consistent way of defining the delay must be used.

## Appendix B. Contributor Addresses

Clarence Filsfils  
Cisco Systems  
Email: cfilsfil@cisco.com

Siva Sivabalan  
Cisco Systems  
Email: msiva@cisco.com

George Swallow  
Cisco Systems  
Email: swallow@cisco.com

Stefano Previdi  
Cisco Systems, Inc  
Via Del Serafico 200  
Rome 00191  
Italy  
Email: sprevidi@cisco.com

Udayasree Palle  
Huawei Technologies  
Divyashree Techno Park, Whitefield  
Bangalore, Karnataka 560066  
India  
Email: udayasree.palle@huawei.com

Avantika  
Huawei Technologies  
Divyashree Techno Park, Whitefield  
Bangalore, Karnataka 560066  
India  
Email: avantika.sushilkumar@huawei.com

Xian Zhang  
Huawei Technologies  
F3-1-B R&D Center, Huawei Base Bantian, Longgang District  
Shenzhen, Guangdong 518129  
P.R.China  
Email: zhang.xian@huawei.com

Authors' Addresses

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

EMail: dhruv.ietf@gmail.com

Qin Wu  
Huawei Technologies  
101 Software Avenue, Yuhua District  
Nanjing, Jiangsu 210012  
China

EMail: bill.wu@huawei.com

Vishwas Manral  
Ionos Network  
4100 Moorpark Av  
San Jose, CA  
USA

EMail: vishwas.ietf@gmail.com

Zafar Ali  
Cisco Systems

EMail: zali@cisco.com

Kenji Kumaki  
KDDI Corporation

EMail: ke-kumaki@kddi.com

PCE Working Group  
Internet-Draft  
Intended status: Standards Track  
Expires: December 21, 2017

E. Crabbe  
Oracle  
I. Minei  
Google, Inc.  
J. Medved  
Cisco Systems, Inc.  
R. Varga  
Pantheon Technologies SRO  
June 19, 2017

PCEP Extensions for Stateful PCE  
draft-ietf-pce-stateful-pce-21

Abstract

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

Although PCEP explicitly makes no assumptions regarding the information available to the PCE, it also makes no provisions for PCE control of timing and sequence of path computations within and across PCEP sessions. This document describes a set of extensions to PCEP to enable stateful control of MPLS-TE and GMPLS LSPs via PCEP.

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 21, 2017.

Copyright Notice

Copyright (c) 2017 IETF Trust and the persons identified as the document authors. All rights reserved.

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

## Table of Contents

1. Introduction . . . . .	3
1.1. Requirements Language . . . . .	4
2. Terminology . . . . .	4
3. Motivation and Objectives for Stateful PCE . . . . .	5
3.1. Motivation . . . . .	5
3.1.1. Background . . . . .	5
3.1.2. Why a Stateful PCE? . . . . .	6
3.1.3. Protocol vs. Configuration . . . . .	7
3.2. Objectives . . . . .	7
4. New Functions to Support Stateful PCEs . . . . .	8
5. Overview of Protocol Extensions . . . . .	9
5.1. LSP State Ownership . . . . .	9
5.2. New Messages . . . . .	9
5.3. Error Reporting . . . . .	10
5.4. Capability Advertisement . . . . .	10
5.5. IGP Extensions for Stateful PCE Capabilities Advertisement . . . . .	11
5.6. State Synchronization . . . . .	12
5.7. LSP Delegation . . . . .	15
5.7.1. Delegating an LSP . . . . .	15
5.7.2. Revoking a Delegation . . . . .	16
5.7.3. Returning a Delegation . . . . .	18
5.7.4. Redundant Stateful PCEs . . . . .	18
5.7.5. Redefinition on PCE Failure . . . . .	19
5.8. LSP Operations . . . . .	19
5.8.1. Passive Stateful PCE Path Computation Request/Response . . . . .	19
5.8.2. Switching from Passive Stateful to Active Stateful .	21
5.8.3. Active Stateful PCE LSP Update . . . . .	22
5.9. LSP Protection . . . . .	23
5.10. PCEP Sessions . . . . .	23
6. PCEP Messages . . . . .	23
6.1. The PCRpt Message . . . . .	24
6.2. The PCUpd Message . . . . .	26
6.3. The PCErr Message . . . . .	28
6.4. The PCReq Message . . . . .	29

6.5.	The PCRep Message . . . . .	30
7.	Object Formats . . . . .	30
7.1.	OPEN Object . . . . .	30
7.1.1.	Stateful PCE Capability TLV . . . . .	30
7.2.	SRP Object . . . . .	31
7.3.	LSP Object . . . . .	33
7.3.1.	LSP-IDENTIFIERS TLVs . . . . .	35
7.3.2.	Symbolic Path Name TLV . . . . .	38
7.3.3.	LSP Error Code TLV . . . . .	39
7.3.4.	RSVP Error Spec TLV . . . . .	40
8.	IANA Considerations . . . . .	41
8.1.	PCE Capabilities in IGP Advertisements . . . . .	41
8.2.	PCEP Messages . . . . .	41
8.3.	PCEP Objects . . . . .	42
8.4.	LSP Object . . . . .	42
8.5.	PCEP-Error Object . . . . .	43
8.6.	Notification Object . . . . .	43
8.7.	PCEP TLV Type Indicators . . . . .	44
8.8.	STATEFUL-PCE-CAPABILITY TLV . . . . .	44
8.9.	LSP-ERROR-CODE TLV . . . . .	45
9.	Manageability Considerations . . . . .	45
9.1.	Control Function and Policy . . . . .	45
9.2.	Information and Data Models . . . . .	46
9.3.	Liveness Detection and Monitoring . . . . .	47
9.4.	Verifying Correct Operation . . . . .	47
9.5.	Requirements on Other Protocols and Functional Components . . . . .	47
9.6.	Impact on Network Operation . . . . .	47
10.	Security Considerations . . . . .	48
10.1.	Vulnerability . . . . .	48
10.2.	LSP State Snooping . . . . .	48
10.3.	Malicious PCE . . . . .	49
10.4.	Malicious PCC . . . . .	49
11.	Contributing Authors . . . . .	49
12.	Acknowledgements . . . . .	50
13.	References . . . . .	50
13.1.	Normative References . . . . .	50
13.2.	Informative References . . . . .	51
	Authors' Addresses . . . . .	53

## 1. Introduction

[RFC5440] describes the Path Computation Element Communication Protocol (PCEP). PCEP defines the communication between a Path Computation Client (PCC) and a Path Computation Element (PCE), or between PCEs, enabling computation of Multiprotocol Label Switching (MPLS) for Traffic Engineering Label Switched Path (TE LSP) characteristics. Extensions for support of Generalized MPLS (GMPLS) in PCEP are defined in [I-D.ietf-pce-gmpls-pcep-extensions]

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

Extensions to permit the PCE to drive creation of an LSP are defined in [I-D.ietf-pce-pce-initiated-lsp], which specifies PCE-initiated LSP creation.

### 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

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

This document uses the following terms defined in [RFC4655]: TED.

This document uses the following terms defined in [RFC3031]: LSP.

This document uses the following terms defined in [RFC8051]: Stateful PCE, Passive Stateful PCE, Active Stateful PCE, Delegation, LSP State Database.

The following terms are defined in this document:

**Revocation:** an operation performed by a PCC on a previously delegated LSP. Revocation revokes the rights granted to the PCE in the delegation operation.

**Redelegation Timeout Interval:** the period of time a PCC waits for, when a PCEP session is terminated, before revoking LSP delegation to a PCE and attempting to redelegate LSPs associated with the terminated PCEP session to an alternate PCE. The Redelegation Timeout Interval is a PCC-local value that can be either operator-configured or dynamically computed by the PCC based on local policy.

**State Timeout Interval:** the period of time a PCC waits for, when a PCEP session is terminated, before flushing LSP state associated with that PCEP session and reverting to operator-defined default parameters or behaviors. The State Timeout Interval is a PCC-

local value that can be either operator-configured or dynamically computed by the PCC based on local policy.

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

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

SRP-ID-number: a number used to correlate errors and LSP State Reports to LSP Update Requests. It is carried in the SRP (Stateful PCE Request Parameters) Object described in Section 7.2.

Within this document, PCEP communications are described through PCC-PCE relationship. The PCE architecture also supports the PCE-PCE communication, by having the requesting PCE fill the role of a PCC, as usual.

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

### 3. Motivation and Objectives for Stateful PCE

#### 3.1. Motivation

[RFC8051] presents several use cases, demonstrating scenarios that benefit from the deployment of a stateful PCE. The scenarios apply equally to MPLS-TE and GMPLS deployments.

##### 3.1.1. Background

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

Unfortunately, this visibility in terms of global LSP state may result in a number of issues for some demand patterns, particularly within a common setup and hold priority. This issue affects online traffic engineering systems.



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

### 3.1.2. Why a Stateful PCE?

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

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

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

Despite these concerns in terms of implementation complexity and scalability, several TE algorithms exist today that have been demonstrated to be extremely effective in large TE systems, providing both rapid convergence and significant benefits in terms of optimality of resource usage [MXMN-TE]. All of these systems share at least two common characteristics: the requirement for both global visibility of a flow (or in this case, a TE LSP) state and for ordered control of path reservations across devices within the system being controlled. While some approaches have been suggested in order to remove the requirements for ordered control (See [MPLS-PC]), these approaches are highly dependent on traffic distribution, and do not allow for multiple simultaneous LSP priorities representing diffserv classes.

The use cases described in [RFC8051] demonstrate a need for visibility into global inter-PCC LSP state in PCE path computations, and for PCE control of sequence and timing in altering LSP path characteristics within and across PCEP sessions.

### 3.1.3. Protocol vs. Configuration

Note that existing configuration tools and protocols can be used to set LSP state, such as a Command Line Interface (CLI) tool. However, this solution has several shortcomings:

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

### 3.2. Objectives

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

- o Allow a single PCC to interact with a mix of stateless and stateful PCEs simultaneously using the same protocol, i.e. PCEP.
- o Support efficient LSP state synchronization between the PCC and one or more active or passive stateful PCEs.
- o Allow a PCC to delegate control of its LSPs to an active stateful PCE such that a given LSP is under the control of a single PCE at any given time.

- \* A PCC may revoke this delegation at any time during the lifetime of the LSP. If LSP delegation is revoked while the PCEP session is up, the PCC MUST notify the PCE about the revocation.
- \* A PCE may return an LSP delegation at any point during the lifetime of the PCEP session. If LSP delegation is returned by the PCE while the PCEP session is up, the PCE MUST notify the PCC about the returned delegation.
- o Allow a PCE to control computation timing and update timing across all LSPs that have been delegated to it.
- o Enable uninterrupted operation of PCC's LSPs in the event of a PCE failure or while control of LSPs is being transferred between PCEs.

#### 4. New Functions to Support Stateful PCEs

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

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

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

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

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

LSP control delegation (C-E,E-C): a PCC grants to a PCE the right to update LSP attributes on one or more LSPs; the PCE becomes the authoritative source of the LSP's attributes as long as the delegation is in effect (See Section 5.7); the PCC may withdraw the delegation or the PCE may give up the delegation at any time.

Similarly to [RFC5440], no assumption is made about the discovery method used by a PCC to discover a set of PCEs (e.g., via static configuration or dynamic discovery) and on the algorithm used to select a PCE.

## 5. Overview of Protocol Extensions

### 5.1. LSP State Ownership

In PCEP (defined in [RFC5440]), LSP state and operation are under the control of a PCC (a PCC may be an LSR or a management station). Attributes received from a PCE are subject to PCC's local policy. The PCEP extensions described in this document do not change this behavior.

An active stateful PCE may have control of a PCC's LSPs that were delegated to it, but the LSP state ownership is retained by the PCC. In particular, in addition to specifying values for LSP's attributes, an active stateful PCE also decides when to make LSP modifications.

Retaining LSP state ownership on the PCC allows for:

- o a PCC to interact with both stateless and stateful PCEs at the same time
- o a stateful PCE to only modify a small subset of LSP parameters, i.e. to set only a small subset of the overall LSP state; other parameters may be set by the operator, for example through command line interface (CLI) commands
- o a PCC to revert delegated LSP to an operator-defined default or to delegate the LSPs to a different PCE, if the PCC get disconnected from a PCE with currently delegated LSPs

### 5.2. New Messages

In this document, we define the following new PCEP messages:

Path Computation State Report (PCRpt): a PCEP message sent by a PCC to a PCE to report the status of one or more LSPs. Each LSP State Report in a PCRpt message MAY contain the actual LSP's path, bandwidth, operational and administrative status, etc. An LSP Status Report carried on a PCRpt message is also used in delegation or revocation of control of an LSP to/from a PCE. The PCRpt message is described in Section 6.1.

Path Computation Update Request (PCUpd): a PCEP message sent by a PCE to a PCC to update LSP parameters, on one or more LSPs. Each LSP Update Request on a PCUpd message MUST contain all LSP parameters that a PCE wishes to be set for a given LSP. An LSP Update Request carried on a PCUpd message is also used to return LSP delegations if at any point PCE no longer desires control of an LSP. The PCUpd message is described in Section 6.2.

The new functions defined in Section 4 are mapped onto the new messages as shown in the following table.

Function	Message
Capability Advertisement (E-C,C-E)	Open
State Synchronization (C-E)	PCRpt
LSP State Report (C-E)	PCRpt
LSP Control Delegation (C-E,E-C)	PCRpt, PCUpd
LSP Update Request (E-C)	PCUpd

Table 1: New Function to Message Mapping

### 5.3. Error Reporting

Error reporting is done using the procedures defined in [RFC5440], and reusing the applicable error types and error values of [RFC5440] wherever appropriate. The current document defines new error values for several error types to cover failures specific to stateful PCE.

### 5.4. Capability Advertisement

During PCEP Initialization Phase, PCEP Speakers (PCE or PCC) advertise their support of stateful PCEP extensions. A PCEP Speaker includes the "Stateful PCE Capability" TLV, described in Section 7.1.1, in the OPEN Object to advertise its support for PCEP stateful extensions. The Stateful Capability TLV includes the 'LSP Update' Flag that indicates whether the PCEP Speaker supports LSP parameter updates.

The presence of the Stateful PCE Capability TLV in PCC's OPEN Object indicates that the PCC is willing to send LSP State Reports whenever LSP parameters or operational status changes.

The presence of the Stateful PCE Capability TLV in PCE's OPEN message indicates that the PCE is interested in receiving LSP State Reports whenever LSP parameters or operational status changes.

The PCEP extensions for stateful PCEs MUST NOT be used if one or both PCEP Speakers have not included the Stateful PCE Capability TLV in their respective OPEN message. If the PCEP Speaker on the PCC supports the extensions of this draft but did not advertise this capability, then upon receipt of PCUpd message from the PCE, it MUST generate a PCErr with error-type 19 (Invalid Operation), error-value 2 (Attempted LSP Update Request if the stateful PCE capability was not advertised)(see Section 8.5) and it SHOULD terminate the PCEP

session. If the PCEP Speaker on the PCE supports the extensions of this draft but did not advertise this capability, then upon receipt of a PCRpt message from the PCC, it MUST generate a PCErr with error-type 19 (Invalid Operation), error-value 5 (Attempted LSP State Report if stateful PCE capability was not advertised) (see Section 8.5) and it SHOULD terminate the PCEP session.

LSP delegation and LSP update operations defined in this document may only be used if both PCEP Speakers set the LSP-UPDATE-CAPABILITY Flag in the "Stateful Capability" TLV to 'Updates Allowed (U Flag = 1)'. If this is not the case and LSP delegation or LSP update operations are attempted, then a PCErr with error-type 19 (Invalid Operation) and error-value 1 (Attempted LSP Update Request for a non-delegated LSP) (see Section 8.5) MUST be generated. Note that, even if one of the PCEP speakers does not set the LSP-UPDATE-CAPABILITY flag in its "Stateful Capability" TLV, a PCE can still operate as a passive stateful PCE by accepting LSP State Reports from the PCC in order to build and maintain an up to date view of the state of the PCC's LSPs.

#### 5.5. IGP Extensions for Stateful PCE Capabilities Advertisement

When PCCs are LSRs participating in the IGP (OSPF or IS-IS), and PCEs are either LSRs or servers also participating in the IGP, an effective mechanism for PCE discovery within an IGP routing domain consists of utilizing IGP advertisements. Extensions for the advertisement of PCE Discovery Information are defined for OSPF and for IS-IS in [RFC5088] and [RFC5089] respectively.

The PCE-CAP-FLAGS sub-TLV, defined in [RFC5089], is an optional sub-TLV used to advertise PCE capabilities. It MAY be present within the PCED sub-TLV carried by OSPF or IS-IS. [RFC5088] and [RFC5089] provide the description and processing rules for this sub-TLV when carried within OSPF and IS-IS, respectively.

The format of the PCE-CAP-FLAGS sub-TLV is included below for easy reference:

Type: 5

Length: Multiple of 4.

Value: This contains an array of units of 32 bit flags with the most significant bit as 0. Each bit represents one PCE capability.

PCE capability bits are defined in [RFC5088]. This document defines new capability bits for the stateful PCE as follows:

Bit	Capability
11	Active Stateful PCE capability
12	Passive Stateful PCE capability

Note that while active and passive stateful PCE capabilities may be advertised during discovery, PCEP Speakers that wish to use stateful PCEP MUST negotiate stateful PCEP capabilities during PCEP session setup, as specified in the current document. A PCC MAY initiate stateful PCEP capability negotiation at PCEP session setup even if it did not receive any IGP PCE capability advertisements.

## 5.6. State Synchronization

The purpose of State Synchronization is to provide a checkpoint-in-time state replica of a PCC's LSP state in a PCE. State Synchronization is performed immediately after the Initialization phase ([RFC5440]).

During State Synchronization, a PCC first takes a snapshot of the state of its LSPs state, then sends the snapshot to a PCE in a sequence of LSP State Reports. Each LSP State Report sent during State Synchronization has the SYNC Flag in the LSP Object set to 1. The set of LSPs for which state is synchronized with a PCE is determined by the PCC's local configuration (see more details in Section 9.1) and MAY also be determined by stateful PCEP capabilities defined in other documents, such as [I-D.ietf-pce-stateful-sync-optimizations].

The end of synchronization marker is a PCRpt message with the SYNC Flag set to 0 for an LSP Object with PLSP-ID equal to the reserved value 0 (see Section 7.3). In this case, the LSP Object SHOULD NOT include the SYMBOLIC-PATH-NAME TLV and SHOULD include the LSP-IDENTIFIERS TLV with the special value of all zeroes. The PCRpt message MUST include an empty ERO as its intended path and SHOULD NOT include the optional RRO object for its actual path. If the PCC has no state to synchronize, it SHOULD only send the end of synchronization marker.

A PCE SHOULD NOT send PCUpd messages to a PCC before State Synchronization is complete. A PCC SHOULD NOT send PCReq messages to a PCE before State Synchronization is complete. This is to allow the PCE to get the best possible view of the network before it starts computing new paths.

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 reestablishment 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 LSP state synchronization, it MUST send a PCErr message with error-type 20 (LSP State Synchronization Error) and error-value 5 (indicating an internal PCC error) to the PCE and terminate the session.

The PCE does not send positive acknowledgements for properly received synchronization messages. It MUST respond with a PCErr message with error-type 20 (LSP State Synchronization Error) and error-value 1 (indicating an error in processing the PCRpt) (see Section 8.5) if it encounters a problem with the LSP State Report it received from the PCC and it MUST terminate the session.

A PCE implementing a limit on the resources a single PCC can occupy, MUST send a PCNtf message with Notification Type 4 (Stateful PCE resource limit exceeded) and Notification Value 1 (Entering resource limit exceeded state) in response to the PCRpt message triggering this condition in the synchronization phase and MUST terminate the session.

The successful State Synchronization sequence is shown in Figure 1.

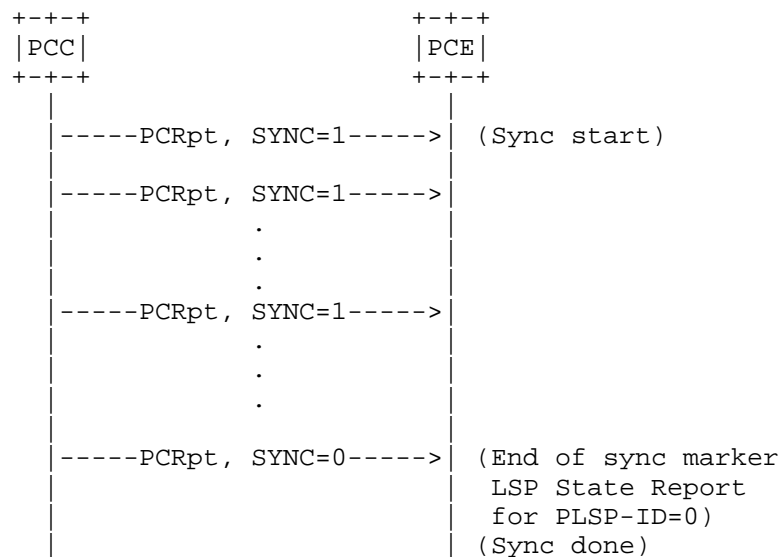


Figure 1: Successful state synchronization



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

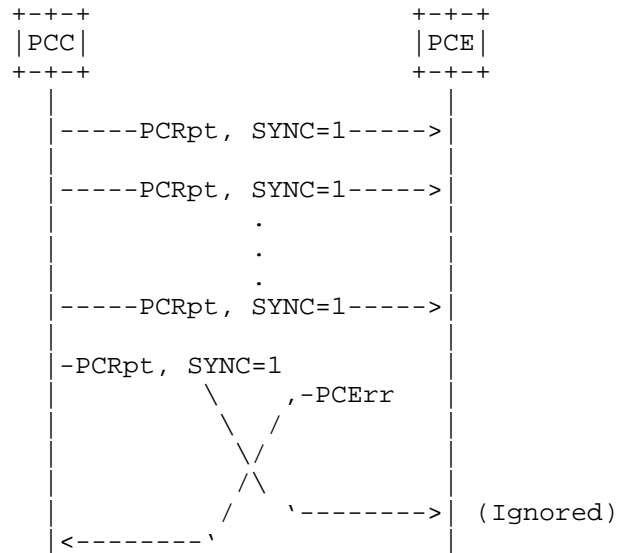


Figure 2: Failed state synchronization (PCE failure)

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

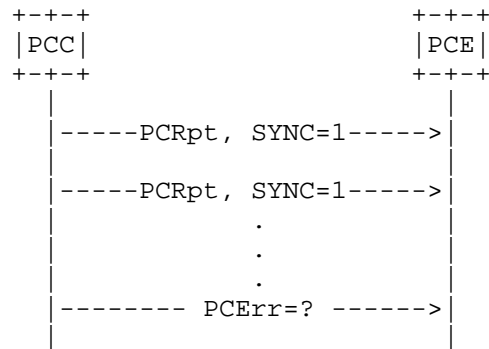


Figure 3: Failed state synchronization (PCC failure)

Optimizations to the synchronization procedures and alternate mechanisms of providing the synchronization function are outside the scope of this document and are discussed elsewhere (see [I-D.ietf-pce-stateful-sync-optimizations]).

### 5.7. LSP Delegation

If during Capability advertisement both the PCE and the PCC have indicated that they support LSP Update, then the PCC may choose to grant the PCE a temporary right to update (a subset of) LSP attributes on one or more LSPs. This is called "LSP Delegation", and it MAY be performed at any time after the Initialization phase, including during the State Synchronization phase.

A PCE MAY return an LSP delegation at any time if it no longer wishes to update the LSP's state. A PCC MAY revoke an LSP delegation at any time. Delegation, Revocation, and Return are done individually for each LSP.

In the event of a delegation being rejected or returned by a PCE, the PCC SHOULD react based on local policy. It can, for example, either retry delegating to the same PCE using an exponentially increasing timer or delegate to an alternate PCE.

#### 5.7.1. Delegating an LSP

A PCC delegates an LSP to a PCE by setting the Delegate flag in LSP State Report to 1. If the PCE does not accept the LSP Delegation, it MUST immediately respond with an empty LSP Update Request which has the Delegate flag set to 0. If the PCE accepts the LSP Delegation, it MUST set the Delegate flag to 1 when it sends an LSP Update Request for the delegated LSP (note that this may occur at a later time). The PCE MAY also immediately acknowledge a delegation by sending an empty LSP Update Request which has the Delegate flag set to 1.

The delegation sequence is shown in Figure 4.

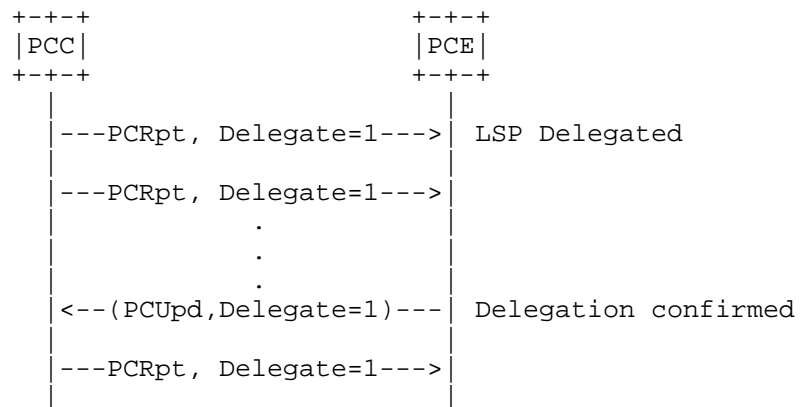


Figure 4: Delegating an LSP

Note that for an LSP to remain delegated to a PCE, the PCC MUST set the Delegate flag to 1 on each LSP State Report sent to the PCE.

## 5.7.2. Revoking a Delegation

### 5.7.2.1. Explicit Revocation

When a PCC decides that a PCE is no longer permitted to modify an LSP, it revokes that LSP's delegation to the PCE. A PCC may revoke an LSP delegation at any time during the LSP's life time. A PCC revoking an LSP delegation MAY immediately remove the updated parameters provided by the PCE and revert to the operator-defined parameters, but to avoid traffic loss, it SHOULD do so in a make-before-break fashion. If the PCC has received but not yet acted on PCUpd messages from the PCE for the LSP whose delegation is being revoked, then it SHOULD ignore these PCUpd messages when processing the message queue. All effects of all messages for which processing started before the revocation took place MUST be allowed to complete and the result MUST be given the same treatment as any LSP that had been previously delegated to the PCE (e.g. the state MAY immediately revert to the operator-defined parameters).

If a PCEP session with the PCE to which the LSP is delegated exists in the UP state during the revocation, the PCC MUST notify that PCE by sending an LSP State Report with the Delegate flag set to 0, as shown in Figure 5.

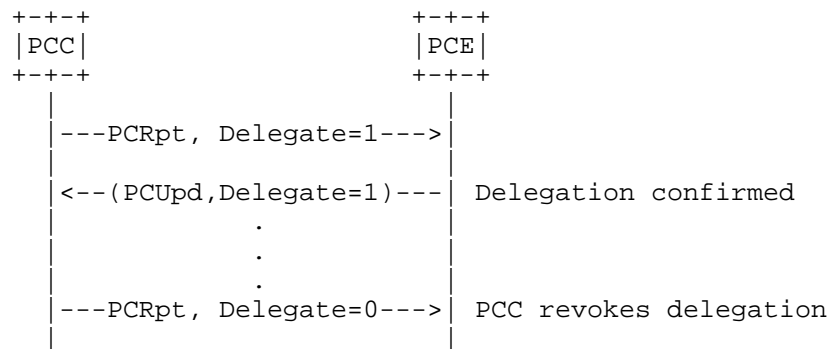


Figure 5: Revoking a Delegation

After an LSP delegation has been revoked, a PCE can no longer update LSP's parameters; an attempt to update parameters of a non-delegated LSP will result in the PCC sending a PCErr message with error-type 19 (Invalid Operation), error-value 1 (attempted LSP Update Request for a non-delegated LSP) (see Section 8.5).

#### 5.7.2.2. Revocation on Redelegating Timeout

When a PCC's PCEP session with a PCE terminates unexpectedly, the PCC MUST wait the time interval specified in Redelegating Timeout Interval before revoking LSP delegations to that PCE and attempting to redelegate LSPs to an alternate PCE. If a PCEP session with the original PCE can be reestablished before the Redelegating Timeout Interval timer expires, LSP delegations to the PCE remain intact.

Likewise, when a PCC's PCEP session with a PCE terminates unexpectedly, and the PCC does not succeed in redelegating its LSPs, the PCC MUST wait for the State Timeout Interval before flushing any LSP state associated with that PCE. Note that the State Timeout Interval timer may expire before the PCC has redelegated the LSPs to another PCE, for example if a PCC is not connected to any active stateful PCE or if no connected active stateful PCE accepts the delegation. In this case, the PCC MUST flush any LSP state set by the PCE upon expiration of the State Timeout Interval and revert to operator-defined default parameters or behaviors. This operation SHOULD be done in a make-before-break fashion.

The State Timeout Interval MUST be greater than or equal to the Redelegating Timeout Interval and MAY be set to infinity (meaning that until the PCC specifically takes action to change the parameters set by the PCE, they will remain intact).

### 5.7.3. Returning a Delegation

In order to keep a delegation, a PCE MUST set the Delegate flag to 1 on each LSP Update Request sent to the PCC. A PCE that no longer wishes to update an LSP's parameters SHOULD return the LSP delegation back to the PCC by sending an empty LSP Update Request which has the Delegate flag set to 0. If a PCC receives an LSP Update Request with the Delegate flag set to 0 (whether the LSP Update Request is empty or not), it MUST treat this as a delegation return.

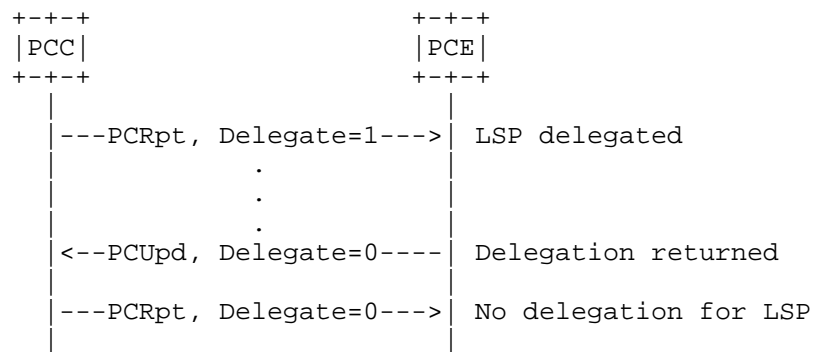


Figure 6: Returning a Delegation

If a PCC cannot delegate an LSP to a PCE (for example, if a PCC is not connected to any active stateful PCE or if no connected active stateful PCE accepts the delegation), the LSP delegation on the PCC will time out within a configurable Redelegating Timeout Interval and the PCC MUST flush any LSP state set by a PCE at the expiration of the State Timeout Interval and revert to operator-defined default parameters or behaviors.

### 5.7.4. Redundant Stateful PCEs

In a redundant configuration where one PCE is backing up another PCE, the backup PCE may have only a subset of the LSPs in the network delegated to it. The backup PCE does not update any LSPs that are not delegated to it. In order to allow the backup to operate in a hot-standby mode and avoid the need for state synchronization in case the primary fails, the backup receives all LSP State Reports from a PCC. When the primary PCE for a given LSP set fails, after expiry of the Redelegating Timeout Interval, the PCC SHOULD delegate to the redundant PCE all LSPs that had been previously delegated to the failed PCE. Assuming that the State Timeout Interval had been configured to be greater than the Redelegating Timeout Interval (as MANDATORY), and assuming that the primary and redundant PCEs take

similar decisions, this delegation change will not cause any changes to the LSP parameters.

#### 5.7.5. Redelegation on PCE Failure

On failure, the goal is to: 1) avoid any traffic loss on the LSPs that were updated by the PCE that crashed 2) minimize the churn in the network in terms of ownership of the LSPs, 3) not leave any "orphan" (undelegated) LSPs and 4) be able to control when the state that was set by the PCE can be changed or purged. The values chosen for the Redelegation Timeout and State Timeout values affect the ability to accomplish these goals.

This section summarizes the behaviour with regards to LSP delegation and LSP state on a PCE failure.

If the PCE crashes but recovers within the Redelegation Timeout, both the delegation state and the LSP state are kept intact.

If the PCE crashes but does not recover within the Redelegation Timeout, the delegation state is returned to the PCC. If the PCC can redelegate the LSPs to another PCE, and that PCE accepts the delegations, there will be no change in LSP state. If the PCC cannot redelegate the LSPs to another PCE, then upon expiration of the State Timeout Interval, the state set by the PCE is removed and the LSP reverts to operator-defined parameters, which may cause a change in the LSP state. Note that an operator may choose to use an infinite State Timeout Interval if he wishes to maintain the PCE state indefinitely. Note also that flushing the state should be implemented using make-before-break to avoid traffic loss.

If there is a standby PCE, the Redelegation Timeout may be set to 0 through policy on the PCC, causing the LSPs to be redelegated immediately to the PCC, which can delegate them immediately to the standby PCE. Assuming that the PCC can redelegate the LSP to the standby PCE within the State Timeout Interval, and assuming the standby PCE takes similar decisions as the failed PCE, the LSP state will be kept intact.

#### 5.8. LSP Operations

##### 5.8.1. Passive Stateful PCE Path Computation Request/Response

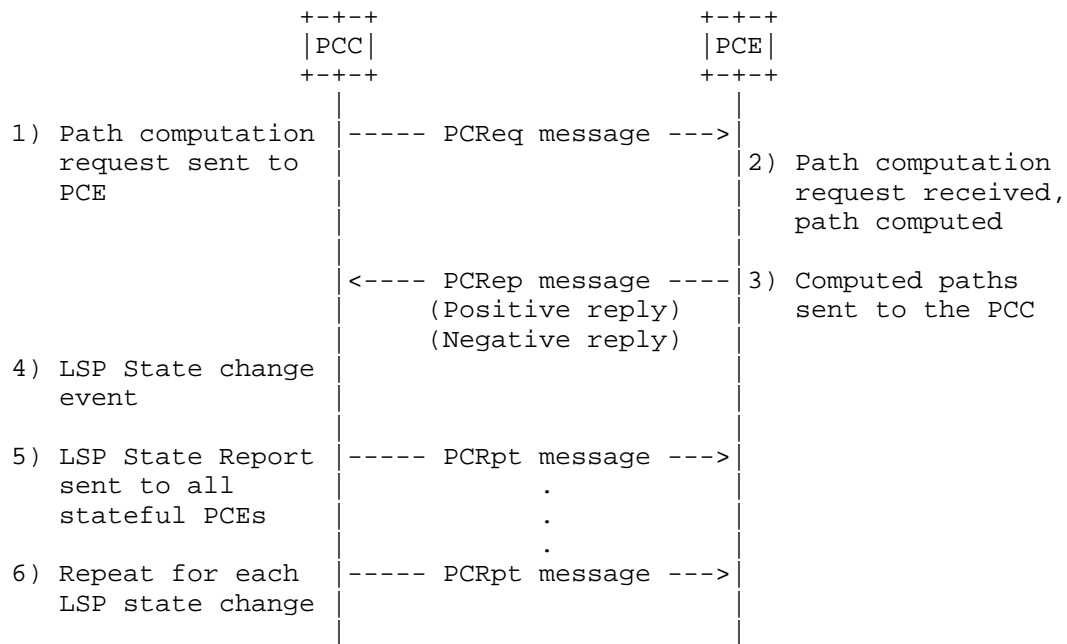


Figure 7: Passive Stateful PCE Path Computation Request/Response

Once a PCC has successfully established a PCEP session with a passive stateful PCE and the PCC's LSP state is synchronized with the PCE (i.e. the PCE knows about all PCC's existing LSPs), if an event is triggered that requires the computation of a set of paths, the PCC sends a path computation request to the PCE ([RFC5440], Section 4.2.3). The PCReq message MAY contain the LSP Object to identify the LSP for which the path computation is requested.

Upon receiving a path computation request from a PCC, the PCE triggers a path computation and returns either a positive or a negative reply to the PCC ([RFC5440], Section 4.2.4).

Upon receiving a positive path computation reply, the PCC receives a set of computed paths and starts to setup the LSPs. For each LSP, it MAY send an LSP State Report carried on a PCRpt message to the PCE, indicating that the LSP's status is "Going-up".

Once an LSP is up or active, the PCC MUST send an LSP State Report carried on a PCRpt message to the PCE, indicating that the LSP's status is 'Up' or 'Active' respectively. If the LSP could not be set up, the PCC MUST send an LSP State Report indicating that the LSP is "Down" and stating the cause of the failure. Note that due to timing constraints, the LSP status may change from 'Going-up' to 'Up' (or

'Down') before the PCC has had a chance to send an LSP State Report indicating that the status is 'Going-up'. In such cases, the PCC MAY choose to only send the PCRpt indicating the latest status ('Active', 'Up' or 'Down').

Upon receiving a negative reply from a PCE, a PCC MAY resend a modified request or take any other appropriate action. For each requested LSP, it SHOULD also send an LSP State Report carried on a PCRpt message to the PCE, indicating that the LSP's status is 'Down'.

There is no direct correlation between PCRep and PCRpt messages. For a given LSP, multiple LSP State Reports will follow a single PCRep message, as a PCC notifies a PCE of the LSP's state changes.

A PCC MUST send each LSP State Report to each stateful PCE that is connected to the PCC.

Note that a single PCRpt message MAY contain multiple LSP State Reports.

The passive stateful model for stateful PCEs is described in [RFC4655], Section 6.8.

#### 5.8.2. Switching from Passive Stateful to Active Stateful

This section deals with the scenario of an LSP transitioning from a passive stateful to an active stateful mode of operation. When the LSP has no working path, prior to delegating the LSP, the PCC MUST first use the procedure defined in Section 5.8.1 to request the initial path from the PCE. This is required because the action of delegating the LSP to a PCE using a PCRpt message is not an explicit request to the PCE to compute a path for the LSP. The only explicit way for a PCC to request a path from PCE is to send a PCReq message. The PCRpt message MUST NOT be used by the PCC to attempt to request a path from the PCE.

When the LSP is delegated after its setup, it may be useful for the PCC to communicate to the PCE the locally configured intended configuration parameters, so that the PCE may reuse them in its computations. Such parameters MAY be acquired through an out of band channel, or MAY be communicated in the PCRpt message delegating the LSPs, by including them as part of the intended-attribute-list as explained in Section 6.1. An implementation MAY allow policies on the PCC to determine the configuration parameters to be sent to the PCE.



## 5.8.3. Active Stateful PCE LSP Update

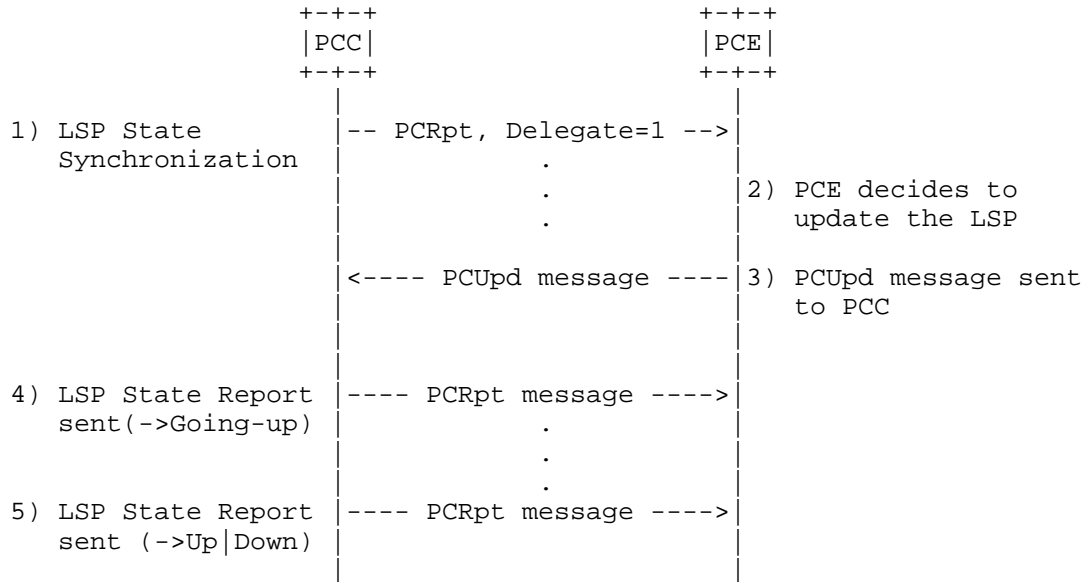


Figure 8: Active Stateful PCE

Once a PCC has successfully established a PCEP session with an active stateful PCE, the PCC's LSP state is synchronized with the PCE (i.e. the PCE knows about all PCC's existing LSPs). After LSPs have been delegated to the PCE, the PCE can modify LSP parameters of delegated LSPs.

To update an LSP, a PCE MUST send the PCC an LSP Update Request using a PCUpd message. The LSP Update Request contains a variety of objects that specify the set of constraints and attributes for the LSP's path. Each LSP Update Request MUST have a unique identifier, the SRP-ID-number, carried in the SRP (Stateful PCE Request Parameters) Object described in Section 7.2. The SRP-ID-number is used to correlate errors and state reports to LSP Update Requests. A single PCUpd message MAY contain multiple LSP Update Requests.

Upon receiving a PCUpd message the PCC starts to setup LSPs specified in LSP Update Requests carried in the message. For each LSP, it MAY send an LSP State Report carried on a PCRpt message to the PCE, indicating that the LSP's status is 'Going-up'. If the PCC decides that the LSP parameters proposed in the PCUpd message are unacceptable, it MUST report this error by including the LSP-ERROR-CODE TLV (Section 7.3.3) with LSP error-value="Unacceptable parameters" in the LSP object in the PCRpt message to the PCE. Based

on local policy, it MAY react further to this error by revoking the delegation. If the PCC receives a PCUpd message for an LSP object identified with a PLSP-ID that does not exist on the PCC, it MUST generate a PCErr with error-type 19 (Invalid Operation), error-value 3, (Attempted LSP Update Request for an LSP identified by an unknown PSP-ID) (see Section 8.5).

Once an LSP is up, the PCC MUST send an LSP State Report (PCRpt message) to the PCE, indicating that the LSP's status is 'Up'. If the LSP could not be set up, the PCC MUST send an LSP State Report indicating that the LSP is 'Down' and stating the cause of the failure. A PCC MAY compress LSP State Reports to only reflect the most up to date state, as discussed in the previous section.

A PCC MUST send each LSP State Report to each stateful PCE that is connected to the PCC.

PCErr and PCRpt messages triggered as a result of a PCUpd message MUST include the SRP-ID-number from the PCUpd. This provides correlation of requests and errors and acknowledgement of state processing. The PCC MAY compress state when processing PCUpd. In this case, receipt of a higher SRP-ID-number implicitly acknowledges processing all the updates with lower SRP-ID-number for the specific LSP (as per Section 7.2).

A PCC MUST NOT send to any PCE a Path Computation Request for a delegated LSP. Should the PCC decide it wants to issue a Path Computation Request on a delegated LSP, it MUST perform Delegation Revocation procedure first.

## 5.9. LSP Protection

LSP protection and interaction with stateful PCE, as well as the extensions necessary to implement this functionality will be discussed in a separate document.

## 5.10. PCEP Sessions

A permanent PCEP session MUST be established between a stateful PCE and the PCC. In the case of session failure, session reestablishment MUST be re-attempted per the procedures defined in [RFC5440].

## 6. PCEP Messages

As defined in [RFC5440], a PCEP message consists of a common header followed by a variable-length body made of a set of objects. For each PCEP message type, a set of rules is defined that specify the set of objects that the message can carry.

### 6.1. The PCRpt Message

A Path Computation LSP State Report message (also referred to as PCRpt message) is a PCEP message sent by a PCC to a PCE to report the current state of an LSP. A PCRpt message can carry more than one LSP State Reports. A PCC can send an LSP State Report either in response to an LSP Update Request from a PCE, or asynchronously when the state of an LSP changes. The Message-Type field of the PCEP common header for the PCRpt message is 10.

The format of the PCRpt message is as follows:

```
<PCRpt Message> ::= <Common Header>
                    <state-report-list>
```

Where:

```
<state-report-list> ::= <state-report>[<state-report-list>]
```

```
<state-report> ::= [<SRP>]
                  <LSP>
                  <path>
```

Where:

```
<path> ::= <intended-path>
          [<actual-attribute-list><actual-path>]
          <intended-attribute-list>
```

```
<actual-attribute-list> ::= [<BANDWIDTH>]
                           [<metric-list>]
```

Where:

```
<intended-path> is represented by the ERO object defined in
section 7.9 of [RFC5440].
<actual-attribute-list> consists of the actual computed and
signaled values of the <BANDWIDTH> and <metric-lists> objects
defined in [RFC5440].
<actual-path> is represented by the RRO object defined in
section 7.10 of [RFC5440].
<intended-attribute-list> is the attribute-list defined in
section 6.5 of [RFC5440] and extended by PCEP extensions.
```

The SRP object (see Section 7.2) is OPTIONAL. If the PCRpt message is not in response to a PCUpd message, the SRP object MAY be omitted. When the PCC does not include the SRP object, the PCE MUST treat this as an SRP object with an SRP-ID-number equal to the reserved value 0x00000000. The reserved value 0x00000000 indicates that the state reported is not as a result of processing a PCUpd message.

If the PCRpt message is in response to a PCUpd message, the SRP object MUST be included and the value of the SRP-ID-number in the SRP Object MUST be the same as that sent in the PCUpd message that triggered the state that is reported. If the PCC compressed several PCUpd messages for the same LSP by only processing the one with the highest number, then it should use the SRP-ID-number of that request. No state compression is allowed for state reporting, e.g. PCRpt messages MUST NOT be pruned from the PCC's egress queue even if subsequent operations on the same LSP have been completed before the PCRpt message has been sent to the TCP stack. The PCC MUST explicitly report state changes (including removal) for paths it manages.

The LSP object (see Section 7.3) is REQUIRED, and it MUST be included in each LSP State Report on the PCRpt message. If the LSP object is missing, the receiving PCE MUST send a PCErr message with Error-type=6 (Mandatory Object missing) and Error-value 8 (LSP object missing).

If the LSP transitioned to non-operational state, the PCC SHOULD include the LSP-ERROR-TLV (Section 7.3.3) with the relevant LSP Error Code to report the error to the PCE.

The intended path, represented by the ERO object, is REQUIRED. If the ERO object is missing, the receiving PCE MUST send a PCErr message with Error-type=6 (Mandatory Object missing) and Error-value 9 (ERO object missing). The ERO may be empty if the PCE does not have a path for a delegated LSP.

The actual path, represented by the RRO object, SHOULD be included in PCRpt by the PCC when the path is up or active, but MAY be omitted if the path is down due to a signaling error or another failure.

The intended-attribute-list maps to the attribute-list in Section 6.5 of [RFC5440] and is used to convey the requested parameters of the LSP path. This is needed in order to support the switch from passive to active stateful PCE as described in Section 5.8.2. When included as part of the intended-attribute-list, the meaning of the BANDWIDTH object is the requested bandwidth as intended by the operator. In this case, the BANDWIDTH Object-Type of 1 SHOULD be used. Similarly, to indicate a limiting constraint, the METRIC object SHOULD be included as part of the intended-attribute-list with the B flag set and with a specific metric value. To indicate the optimization metric, the METRIC object SHOULD be included as part of the intended-attribute-list with the B flag unset and the metric value set to zero. Note that the intended-attribute-list is optional and thus may be omitted. In this case, the PCE MAY use the values in the actual-attribute-list as the requested parameters for the path.

The actual-attribute-list consists of the actual computed and signaled values of the BANDWIDTH and METRIC objects defined in [RFC5440]. When included as part of the actual-attribute-list, Object-Type 2 ([RFC5440]) SHOULD be used for the BANDWIDTH object and the C flag SHOULD be set in the METRIC object ([RFC5440]).

Note that the ordering of intended-path, actual-attribute-list, actual-path and intended-attribute-list is chosen to retain compatibility with implementations of an earlier version of this standard.

A PCE may choose to implement a limit on the resources a single PCC can occupy. If a PCRpt is received that causes the PCE to exceed this limit, the PCE MUST notify the PCC using a PCNtf message with Notification Type 4 (Stateful PCE resource limit exceeded) and Notification Value 1 (Entering resource limit exceeded state) and MUST terminate the session.

## 6.2. The PCUpd Message

A Path Computation LSP Update Request message (also referred to as PCUpd message) is a PCEP message sent by a PCE to a PCC to update attributes of an LSP. A PCUpd message can carry more than one LSP Update Request. The Message-Type field of the PCEP common header for the PCUpd message is 11.

The format of a PCUpd message is as follows:

```
<PCUpd Message> ::= <Common Header>
                        <update-request-list>
```

Where:

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

```
<update-request> ::= <SRP>
                      <LSP>
                      <path>
```

Where:

```
<path> ::= <intended-path><intended-attribute-list>
```

Where:

```
<intended-path> is represented by the ERO object defined in
section 7.9 of [RFC5440].
<intended-attribute-list> is the attribute-list defined in [RFC5440]
and extended by PCEP extensions.
```

There are three mandatory objects that MUST be included within each LSP Update Request in the PCUpd message: the SRP Object (see

Section 7.2), the LSP object (see Section 7.3) and the ERO object (as defined in [RFC5440], which represents the intended path. If the SRP object is missing, the receiving PCC MUST send a PCErr message with Error-type=6 (Mandatory Object missing) and Error-value=10 (SRP object missing). If the LSP object is missing, the receiving PCC MUST send a PCErr message with Error-type=6 (Mandatory Object missing) and Error-value=8 (LSP object missing). If the ERO object is missing, the receiving PCC MUST send a PCErr message with Error-type=6 (Mandatory Object missing) and Error-value=9 (ERO object missing).

The ERO in the PCUpd may be empty if the PCE cannot find a valid path for a delegated LSP. One typical situation resulting in this empty ERO carried in the PCUpd message is that a PCE can no longer find a strict SRLG-disjoint path for a delegated LSP after a link failure. The PCC SHOULD implement a local policy to decide the appropriate action to be taken: either tear down the LSP, or revoke the delegation and use a locally computed path, or keep the existing LSP.

A PCC only acts on an LSP Update Request if permitted by the local policy configured by the network manager. Each LSP Update Request that the PCC acts on results in an LSP setup operation. An LSP Update Request MUST contain all LSP parameters that a PCE wishes to be set for the LSP. A PCC MAY set missing parameters from locally configured defaults. If the LSP specified in the Update Request is already up, it will be re-signaled.

The PCC SHOULD minimize the traffic interruption, and MAY use the make-before-break procedures described in [RFC3209] in order to achieve this goal. If the make-before-break procedures are used, two paths will briefly co-exist. The PCC MUST send separate PCRpt messages for each, identified by the LSP-IDENTIFIERS TLV. When the old path is torn down after the head end switches over the traffic, this event MUST be reported by sending a PCRpt message with the LSP-IDENTIFIERS-TLV of the old path and the R bit set. The SRP-ID-number that the PCC associates with this PCRpt MUST be 0x00000000. Thus, a make-before-break operation will typically result in at least two PCRpt messages, one for the new path and one for the removal of the old path (more messages may be possible if intermediate states are reported).

If the path setup fails due to an RSVP signaling error, the error is reported to the PCE. The PCC will not attempt to resignal the path until it is prompted again by the PCE with a subsequent PCUpd message.

A PCC MUST respond with an LSP State Report to each LSP Update Request it processed to indicate the resulting state of the LSP in

the network (even if this processing did not result in changing the state of the LSP). The SRP-ID-number included in the PCRpt MUST match that in the PCUpd. A PCC MAY respond with multiple LSP State Reports to report LSP setup progress of a single LSP. In that case, the SRP-ID-number MUST be included for the first message, for subsequent messages the reserved value 0x00000000 SHOULD be used.

Note that a PCC MUST process all LSP Update Requests - for example, an LSP Update Request is sent when a PCE returns delegation or puts an LSP into non-operational state. The protocol relies on TCP for message-level flow control.

If the rate of PCUpd messages sent to a PCC for the same target LSP exceeds the rate at which the PCC can signal LSPs into the network, the PCC MAY perform state compression on its ingress queue. The compression algorithm is based on the fact that each PCUpd request contains the complete LSP state the PCE wishes to be set and works as follows: when the PCC starts processing a PCUpd message at the head of its ingress queue, it may search the queue forward for more recent PCUpd messages pertaining that particular LSP, prune all but the latest one from the queue and process only the last one as that request contains the most up-to-date desired state for the LSP. The PCC MUST NOT send PCRpt nor PCErr messages for requests which were pruned from the queue in this way. This compression step may be performed only while the LSP is not being signaled, e.g. if two PCUpd arrive for the same LSP in quick succession and the PCC started the signaling of the changes relevant to the first PCUpd, then it MUST wait until the signaling finishes (and report the new state via a PCRpt) before attempting to apply the changes indicated in the second PCUpd.

Note also that it is up to the PCE to handle inter-LSP dependencies; for example, if ordering of LSP set-ups is required, the PCE has to wait for an LSP State Report for a previous LSP before starting the update of the next LSP.

If the PCUpd cannot be satisfied (for example due to unsupported object or TLV), the PCC MUST respond with a PCErr message indicating the failure (see Section 7.3.3).

### 6.3. The PCErr Message

If the stateful PCE capability has been advertised on the PCEP session, the PCErr message MAY include the SRP object. If the error reported is the result of an LSP update request, then the SRP-ID-number MUST be the one from the PCUpd that triggered the error. If the error is unsolicited, the SRP object MAY be omitted. This is

equivalent to including an SRP object with SRP-ID-number equal to the reserved value 0x00000000.

The format of a PCErr message from [RFC5440] is extended 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> | <stateful-request-id-list>]
           <error-obj-list>

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

<stateful-request-id-list> ::= <SRP> [<stateful-request-id-list>]

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

```

#### 6.4. The PCReq Message

A PCC MAY include the LSP object in the PCReq message (see Section 7.3) if the stateful PCE capability has been negotiated on a PCEP session between the PCC and a PCE.

The definition of the PCReq message from [RFC5440] is extended to optionally include the LSP object after the END-POINTS object. The encoding from [RFC5440] will become:

```

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

```

Where:

```

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

<request> ::= <RP>
              <END-POINTS>
              [<LSP>]
              [<LSPA>]
              [<BANDWIDTH>]
              [<metric-list>]
              [<RRO> [<BANDWIDTH>]]
              [<IRO>]
              [<LOAD-BALANCING>]

```



## 6.5. The PCRep Message

A PCE MAY include the LSP object in the PCRep message (see (Section 7.3) if the stateful PCE capability has been negotiated on a PCEP session between the PCC and the PCE and the LSP object was included in the corresponding PCReq message from the PCC.

The definition of the PCRep message from [RFC5440] is extended to optionally include the LSP object after the RP object. The encoding from [RFC5440] will become:

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

Where:

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

<response> ::= <RP>
                [<LSP>]
                [<NO-PATH>]
                [<attribute-list>]
                [<path-list>]
```

## 7. Object Formats

The PCEP objects defined in this document are compliant with the PCEP object format defined in [RFC5440]. The P flag and the I flag of the PCEP objects defined in the current document MUST be set to 0 on transmission and SHOULD be ignored on receipt since the P and I flags are exclusively related to path computation requests.

### 7.1. OPEN Object

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

#### 7.1.1. Stateful PCE Capability TLV

The STATEFUL-PCE-CAPABILITY TLV is an optional TLV for use in the OPEN Object for stateful PCE capability advertisement. Its format is shown in the following figure:



Figure 9: STATEFUL-PCE-CAPABILITY TLV format

The type (16 bits) of the TLV is 16. The length field is 16 bit-long and has a fixed value of 4.

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

U (LSP-UPDATE-CAPABILITY - 1 bit): if set to 1 by a PCC, the U Flag indicates that the PCC allows modification of LSP parameters; if set to 1 by a PCE, the U Flag indicates that the PCE is capable of updating LSP parameters. The LSP-UPDATE-CAPABILITY Flag must be advertised by both a PCC and a PCE for PCUpd messages to be allowed on a PCEP session.

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

A PCEP speaker operating in passive stateful PCE mode advertises the stateful PCE capability with the U flag set to 0. A PCEP speaker operating in active stateful PCE mode advertises the stateful PCE capability with the U Flag set to 1.

Advertisement of the stateful PCE capability implies support of LSPs that are signaled via RSVP, as well as the objects, TLVs and procedures defined in this document.

## 7.2. SRP Object

The SRP (Stateful PCE Request Parameters) object MUST be carried within PCUpd messages and MAY be carried within PCRpt and PCErr messages. The SRP object is used to correlate between update requests sent by the PCE and the error reports and state reports sent by the PCC.

SRP Object-Class is 33.

SRP Object-Type is 1.

The format of the SRP object body is shown in Figure 10:

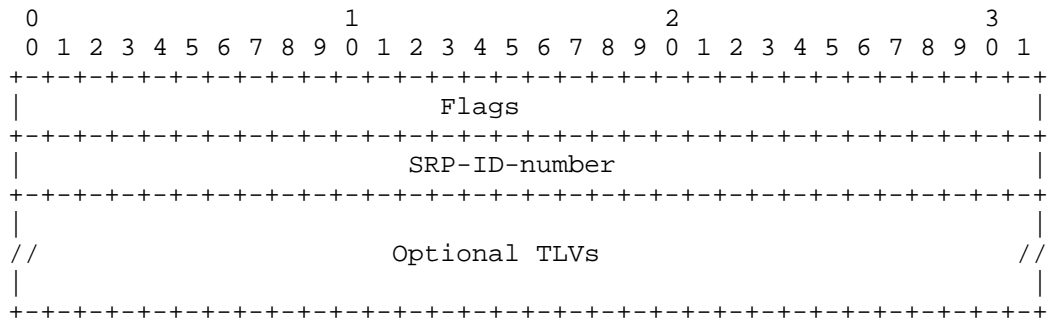


Figure 10: The SRP Object format

The SRP object body has a variable length and may contain additional TLVs.

Flags (32 bits): None defined yet.

SRP-ID-number (32 bits): The SRP-ID-number value in the scope of the current PCEP session uniquely identify the operation that the PCE has requested the PCC to perform on a given LSP. The SRP-ID-number is incremented each time a new request is sent to the PCC, and may wrap around.

The values 0x00000000 and 0xFFFFFFFF are reserved.

Optional TLVs MAY be included within the SRP object body. The specification of such TLVs is outside the scope of this document.

Every request to update an LSP receives a new SRP-ID-number. This number is unique per PCEP session and is incremented each time an operation is requested from the PCE. Thus, for a given LSP there may be more than one SRP-ID-number unacknowledged at a given time. The value of the SRP-ID-number is echoed back by the PCC in PCErr and PCRpt messages to allow for correlation between requests made by the PCE and errors or state reports generated by the PCC. If the error or report were not as a result of a PCE operation (for example in the case of a link down event), the reserved value of 0x00000000 is used for the SRP-ID-number. The absence of the SRP object is equivalent to an SRP object with the reserved value of 0x00000000. An SRP-ID-number is considered unacknowledged and cannot be reused until a PCErr or PCRpt arrives with an SRP-ID-number equal or higher for the same LSP. In case of SRP-ID-number wrapping the last SRP-ID-number before the wrapping MUST be explicitly acknowledged, to avoid a situation where SRP-ID-numbers remain unacknowledged after the wrap.

This means that the PCC may need to issue two PCUpd messages on detecting a wrap.

### 7.3. LSP Object

The LSP object MUST be present within PCRpt and PCUpd messages. The LSP object MAY be carried within PCReq and PCRep messages if the stateful PCE capability has been negotiated on the session. The LSP object contains a set of fields used to specify the target LSP, the operation to be performed on the LSP, and LSP Delegation. It also contains a flag indicating to a PCE that the LSP state synchronization is in progress. This document focuses on LSPs that are signaled with RSVP, many of the TLVs used with the LSP object mirror RSVP state.

LSP Object-Class is 32.

LSP Object-Type is 1.

The format of the LSP object body is shown in Figure 11:

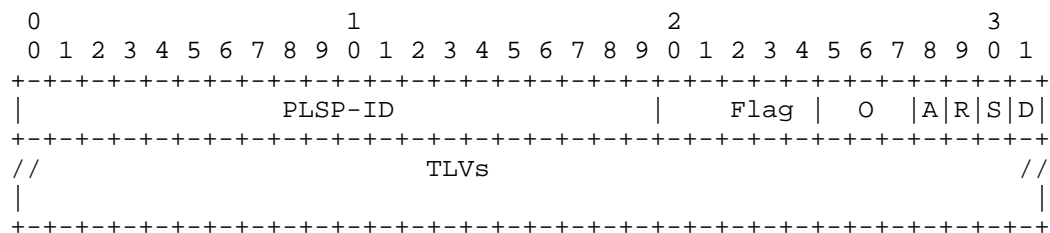


Figure 11: The LSP Object format

**PLSP-ID (20 bits):** A PCEP-specific identifier for the LSP. A PCC creates a unique PLSP-ID for each LSP that is constant for the lifetime of a PCEP session. The PCC will advertise the same PLSP-ID on all PCEP sessions it maintains at a given time. The mapping of the Symbolic Path Name to PLSP-ID is communicated to the PCE by sending a PCRpt message containing the SYMBOLIC-PATH-NAME TLV. All subsequent PCEP messages then address the LSP by the PLSP-ID. The values of 0 and 0xFFFFF are reserved. Note that the PLSP-ID is a value that is constant for the lifetime of the PCEP session, during which time for an RSVP-signaled LSP there might be a different RSVP identifiers (LSP-id, tunnel-id) allocated to it.

**Flags (12 bits), starting from the least significant bit:**

**D (Delegate - 1 bit):** On a PCRpt message, the D Flag set to 1 indicates that the PCC is delegating the LSP to the PCE. On a

PCUpd message, the D flag set to 1 indicates that the PCE is confirming the LSP Delegation. To keep an LSP delegated to the PCE, the PCC must set the D flag to 1 on each PCRpt message for the duration of the delegation - the first PCRpt with the D flag set to 0 revokes the delegation. To keep the delegation, the PCE must set the D flag to 1 on each PCUpd message for the duration of the delegation - the first PCUpd with the D flag set to 0 returns the delegation.

S (SYNC - 1 bit): The S Flag MUST be set to 1 on each PCRpt sent from a PCC during State Synchronization. The S Flag MUST be set to 0 in other messages sent from the PCC. When sending a PCUpd message, the PCE MUST set the S Flag to 0.

R(Remove - 1 bit): On PCRpt messages the R Flag indicates that the LSP has been removed from the PCC and the PCE SHOULD remove all state from its database. Upon receiving an LSP State Report with the R Flag set to 1 for an RSVP-signaled LSP, the PCE SHOULD remove all state for the path identified by the LSP-IDENTIFIERS TLV from its database. When the all-zeros LSP-IDENTIFIERS TLV is used, the PCE SHOULD remove all state for the PLSP-ID from its database. When sending a PCUpd message, the PCE MUST set the R Flag to 0.

A(Administrative - 1 bit): On PCRpt messages, the A Flag indicates the PCC's target operational status for this LSP. On PCUpd messages, the A Flag indicates the LSP status that the PCE desires for this LSP. In both cases, a value of '1' means that the desired operational state is active, and a value of '0' means that the desired operational state is inactive. A PCC ignores the A flag on a PCUpd message unless the operator's policy allows the PCE to control the corresponding LSP's administrative state.

O(Operational - 3 bits): On PCRpt messages, the O Field represents the operational status of the LSP.

The following values are defined:

0 - DOWN: not active.

1 - UP: signalled.

2 - ACTIVE: up and carrying traffic.

3 - GOING-DOWN: LSP is being torn down, resources are being released.

4 - GOING-UP: LSP is being signalled.

5-7 - Reserved: these values are reserved for future use.

Unassigned bits are considered reserved. They MUST be set to 0 on transmission and MUST be ignored on receipt. When sending a PCUpd message, the PCE MUST set the O Field to 0.

TLVs that may be included in the LSP Object are described in the following sections. Other optional TLVs, that are not defined in this document, MAY also be included within the LSP Object body.

#### 7.3.1. LSP-IDENTIFIERS TLVs

The LSP-IDENTIFIERS TLV MUST be included in the LSP object in PCRpt messages for RSVP-signaled LSPs. If the TLV is missing, the PCE will generate an error with error-type 6 (mandatory object missing) and error-value 11 (LSP-IDENTIFIERS TLV missing) and close the session. The LSP-IDENTIFIERS TLV MAY be included in the LSP object in PCUpd messages for RSVP-signaled LSPs. The special value of all zeros for this TLV is used to refer to all paths pertaining to a particular PLSP-ID. There are two LSP-IDENTIFIERS TLVs, one for IPv4 and one for IPv6.

It is the responsibility of the PCC to send to the PCE the identifiers for each RSVP incarnation of the tunnel. For example, in a make-before-break scenario, the PCC MUST send a separate PCRpt for the old and for the reoptimized paths, and explicitly report removal of any of these paths using the R bit in the LSP object.

The format of the IPV4-LSP-IDENTIFIERS TLV is shown in the following figure:

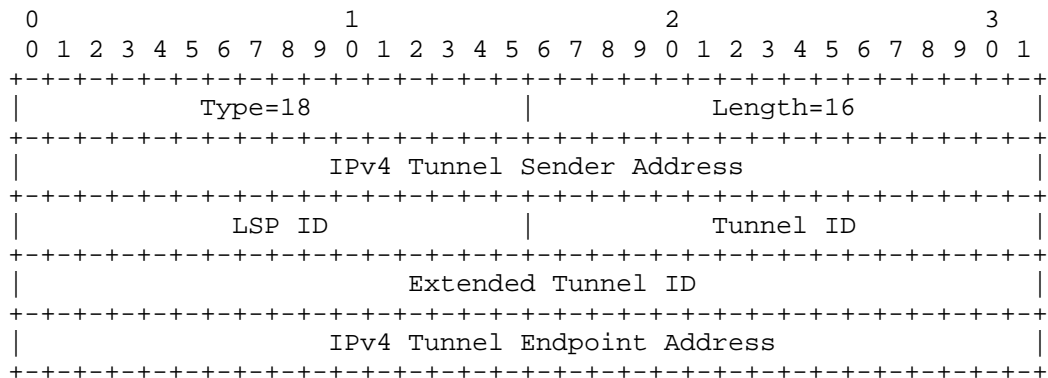


Figure 12: IPV4-LSP-IDENTIFIERS TLV format

The type (16 bits) of the TLV is 18. The length field is 16 bit-long and has a fixed value of 16. The value contains the following fields:

IPv4 Tunnel Sender Address: contains the sender node's IPv4 address, as defined in [RFC3209], Section 4.6.2.1 for the LSP\_TUNNEL\_IPv4 Sender Template Object.

LSP ID: contains the 16-bit 'LSP ID' identifier defined in [RFC3209], Section 4.6.2.1 for the LSP\_TUNNEL\_IPv4 Sender Template Object. A value of 0 MUST be used if the LSP is not yet signaled.

Tunnel ID: contains the 16-bit 'Tunnel ID' identifier defined in [RFC3209], Section 4.6.1.1 for the LSP\_TUNNEL\_IPv4 Session Object.

Extended Tunnel ID: contains the 32-bit 'Extended Tunnel ID' identifier defined in [RFC3209], Section 4.6.1.1 for the LSP\_TUNNEL\_IPv4 Session Object.

IPv4 Tunnel Endpoint Address: contains the egress node's IPv4 address, as defined in [RFC3209], Section 4.6.1.1 for the LSP\_TUNNEL\_IPv4 Sender Template Object.

The format of the IPV6-LSP-IDENTIFIERS TLV is shown in the following figure:

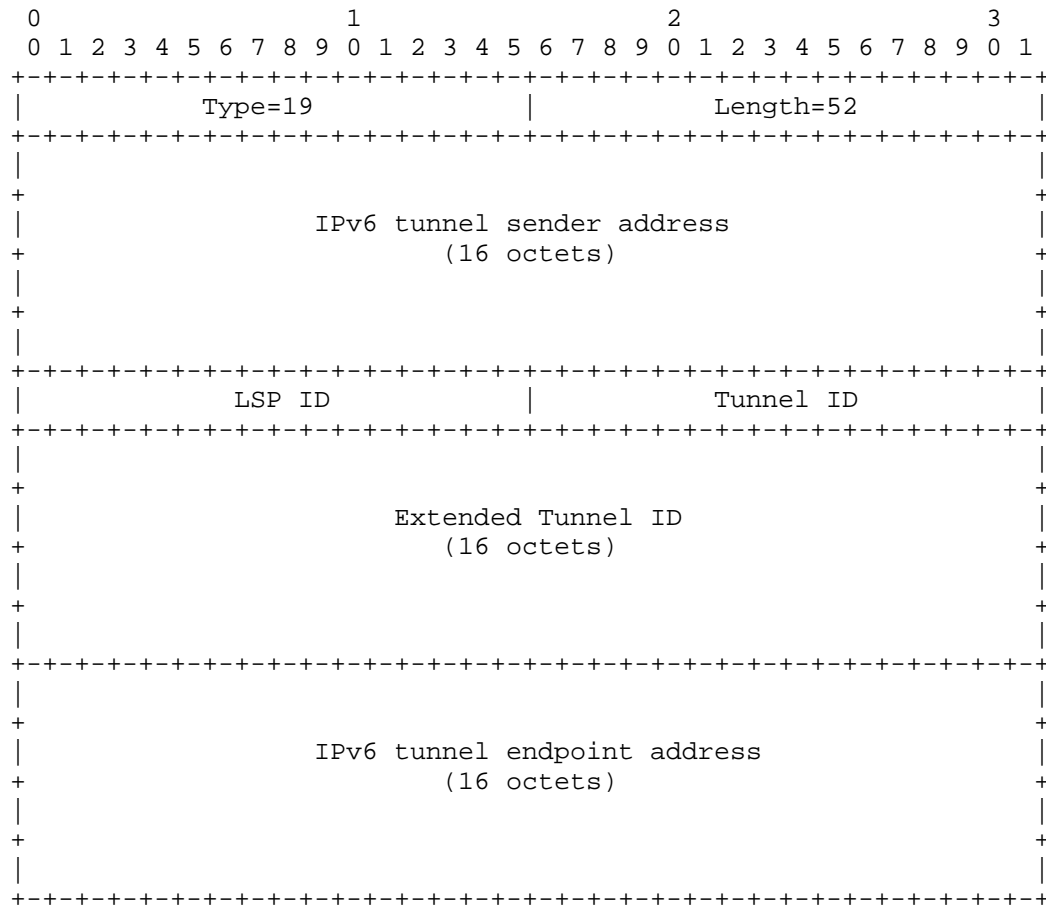


Figure 13: IPV6-LSP-IDENTIFIERS TLV format

The type (16 bits) of the TLV is 19. The length field is 16 bit-long and has a fixed value of 52. The value contains the following fields:

**IPv6 Tunnel Sender Address:** contains the sender node's IPv6 address, as defined in [RFC3209], Section 4.6.2.2 for the LSP\_TUNNEL\_IPv6 Sender Template Object.

**LSP ID:** contains the 16-bit 'LSP ID' identifier defined in [RFC3209], Section 4.6.2.2 for the LSP\_TUNNEL\_IPv6 Sender Template Object. A value of 0 MUST be used if the LSP is not yet signaled.

**Tunnel ID:** contains the 16-bit 'Tunnel ID' identifier defined in [RFC3209], Section 4.6.1.2 for the LSP\_TUNNEL\_IPv6 Session Object.



Extended Tunnel ID: contains the 128-bit 'Extended Tunnel ID' identifier defined in [RFC3209], Section 4.6.1.2 for the LSP\_TUNNEL\_IPv6 Session Object.

IPv6 Tunnel Endpoint Address: contains the egress node's IPv6 address, as defined in [RFC3209], Section 4.6.1.2 for the LSP\_TUNNEL\_IPv6 Session Object.

The Tunnel ID remains constant over the life time of a tunnel.

### 7.3.2. Symbolic Path Name TLV

Each LSP MUST have a symbolic path name that is unique in the PCC. The symbolic path name is a human-readable string that identifies an LSP in the network. The symbolic path name MUST remain constant throughout an LSP's lifetime, which may span across multiple consecutive PCEP sessions and/or PCC restarts. The symbolic path name MAY be specified by an operator in a PCC's configuration. If the operator does not specify a unique symbolic name for an LSP, then the PCC MUST auto-generate one.

The PCE uses the symbolic path name as a stable identifier for the LSP. If the PCEP session restarts, or the PCC restarts, or the PCC re-delegates the LSP to a different PCE, the symbolic path name for the LSP remains constant and can be used to correlate across the PCEP session instances.

The other protocol identifiers for the LSP cannot reliably be used to identify the LSP across multiple PCEP sessions, for the following reasons.

- o The PLSP-ID is unique only within the scope of a single PCEP session.
- o The LSP-IDENTIFIERS TLV is only guaranteed to be present for LSPs that are signalled with RSVP-TE, and may change during the lifetime of the LSP.

The SYMBOLIC-PATH-NAME TLV MUST be included in the LSP object in the LSP State Report (PCRpt) message when during a given PCEP session an LSP is first reported to a PCE. A PCC sends to a PCE the first LSP State Report either during State Synchronization, or when a new LSP is configured at the PCC.

The initial PCRpt creates a binding between the symbolic path name and the PLSP-ID for the LSP which lasts for the duration of the PCEP session. The PCC MAY omit the symbolic path name from subsequent LSP

State Reports for that LSP on that PCEP session, and just use the PLSP-ID.

The format of the SYMBOLIC-PATH-NAME TLV is shown in the following figure:

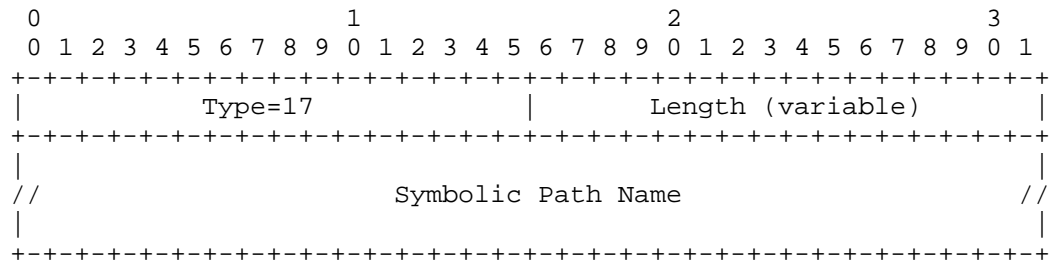


Figure 14: SYMBOLIC-PATH-NAME TLV format

Type (16 bits): The type is 17.

Length (16 bits): indicates the total length of the TLV in octets and MUST be greater than 0. The TLV MUST be zero-padded so that the TLV is 4-octet aligned.

Symbolic Path Name (variable): symbolic name for the LSP, unique in the PCC. It SHOULD be a string of printable ASCII characters, without a NULL terminator.

### 7.3.3. LSP Error Code TLV

The LSP Error code TLV is an optional TLV for use in the LSP object to convey error information. When an LSP Update Request fails, an LSP State Report MUST be sent to report the current state of the LSP, and SHOULD contain the LSP-ERROR-CODE TLV indicating the reason for the failure. Similarly, when a PCRpt is sent as a result of an LSP transitioning to non-operational state, the LSP-ERROR-CODE TLV SHOULD be included to indicate the reason for the transition.

The format of the LSP-ERROR-CODE TLV is shown in the following figure:

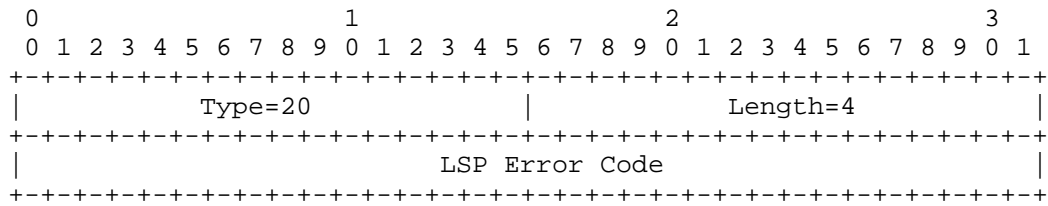


Figure 15: LSP-ERROR-CODE TLV format

The type (16 bits) of the TLV is 20. The length field is 16 bit-long and has a fixed value of 4. The value contains an error code that indicates the cause of the failure.

The following LSP Error Codes are currently defined:

Value	Meaning
1	Unknown reason
2	Limit reached for PCE-controlled LSPs
3	Too many pending LSP update requests
4	Unacceptable parameters
5	Internal error
6	LSP administratively brought down
7	LSP preempted
8	RSVP signaling error

#### 7.3.4. RSVP Error Spec TLV

The RSVP-ERROR-SPEC TLV is an optional TLV for use in the LSP object to carry RSVP error information. It includes the RSVP\_ERROR\_SPEC or USER\_ERROR\_SPEC Object ([RFC2205] and [RFC5284]) which were returned to the PCC from a downstream node. If the set up of an LSP fails at a downstream node which returned an ERROR\_SPEC to the PCC, the PCC SHOULD include in the PCRpt for this LSP the LSP-ERROR-CODE TLV with LSP Error Code = "RSVP signaling error" and the RSVP-ERROR-SPEC TLV with the relevant RSVP\_ERROR\_SPEC or USER\_ERROR\_SPEC Object.

The format of the RSVP-ERROR-SPEC TLV is shown in the following figure:

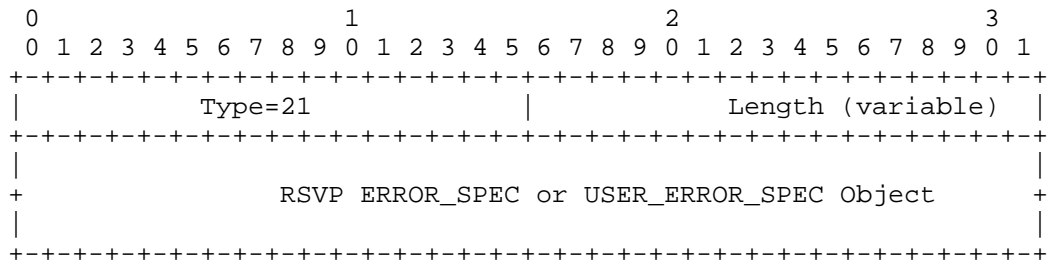


Figure 16: RSVP-ERROR-SPEC TLV format

Type (16 bits): The type is 21.

Length (16 bits): indicates the total length of the TLV in octets. The TLV MUST be zero-padded so that the TLV is 4-octet aligned.

Value (variable): contains the RSVP\_ERROR\_SPEC or USER\_ERROR\_SPEC Object: as specified in [RFC2205] and [RFC5284], including the object header.

## 8. IANA Considerations

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

### 8.1. PCE Capabilities in IGP Advertisements

IANA is requested to confirm the early allocation of the following bits in the OSPF Parameters "PCE Capability Flags" registry, and to update the reference in the registry to point to this document, when it is an RFC:

Bit	Meaning	Reference
11	Active Stateful PCE capability	This document
12	Passive Stateful PCE capability	This document

### 8.2. PCEP Messages

IANA is requested to confirm the early allocation of the following message types within the "PCEP Messages" sub-registry of the PCEP Numbers registry, and to update the reference in the registry to point to this document, when it is an RFC:

Value	Meaning	Reference
10	Report	This document
11	Update	This document

### 8.3. PCEP Objects

IANA is requested to confirm the early allocation of the following object-class values and object types within the "PCEP Objects" sub-registry of the PCEP Numbers registry, and to update the reference in the registry to point to this document, when it is an RFC:

Object-Class Value	Name	Reference
32	LSP Object-Type 1	This document
33	SRP Object-Type 1	This document

### 8.4. LSP Object

This document requests that a new sub-registry, named "LSP 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:

- o Bit number (counting from bit 0 as the most significant bit)
- o Capability description
- o Defining RFC

The following values are defined in this document:

Bit	Description	Reference
0-4	Reserved	This document
5-7	Operational (3 bits)	This document
8	Administrative	This document
9	Remove	This document
10	SYNC	This document
11	Delegate	This document

### 8.5. PCEP-Error Object

IANA is requested to confirm the early allocation of the following Error Types and Error Values within the "PCEP-ERROR Object Error Types and Values" sub-registry of the PCEP Numbers registry, and to update the reference in the registry to point to this document, when it is an RFC:

Error-Type	Meaning
6	Mandatory Object missing
	Error-value=8: LSP Object missing
	Error-value=9: ERO Object missing
	Error-value=10: SRP Object missing
	Error-value=11: LSP-IDENTIFIERS TLV missing
19	Invalid Operation
	Error-value=1: Attempted LSP Update Request for a non-delegated LSP. The PCEP-ERROR Object is followed by the LSP Object that identifies the LSP.
	Error-value=2: Attempted LSP Update Request if the stateful PCE capability was not advertised.
	Error-value=3: Attempted LSP Update Request for an LSP identified by an unknown PLSP-ID.
	Error-value=5: Attempted LSP State Report if stateful PCE capability was not advertised.
20	LSP State synchronization error.
	Error-value=1: A PCE indicates to a PCC that it can not process (an otherwise valid) LSP State Report. The PCEP-ERROR Object is followed by the LSP Object that identifies the LSP.
	Error-value=5: A PCC indicates to a PCE that it can not complete the state synchronization,

### 8.6. Notification Object

IANA is requested to confirm the early allocation of the following Notification Types and Notification Values within the "Notification Object" sub-registry of the PCEP Numbers registry, and to update the reference in the registry to point to this document, when it is an RFC:

Notification-Type	Meaning
4	Stateful PCE resource limit exceeded

Notification-value=1:	Entering resource limit exceeded state
-----------------------	--

Note to IANA: the early allocation included an additional Notification value 2 for "Exiting resource limit exceeded state". This Notification value is no longer required.

### 8.7. PCEP TLV Type Indicators

IANA is requested to confirm the early allocation of the following TLV Type Indicator values within the "PCEP TLV Type Indicators" sub-registry of the PCEP Numbers registry, and to update the reference in the registry to point to this document, when it is an RFC:

Value	Meaning	Reference
16	STATEFUL-PCE-CAPABILITY	This document
17	SYMBOLIC-PATH-NAME	This document
18	IPV4-LSP-IDENTIFIERS	This document
19	IPV6-LSP-IDENTIFIERS	This document
20	LSP-ERROR-CODE	This document
21	RSVP-ERROR-SPEC	This document

### 8.8. STATEFUL-PCE-CAPABILITY TLV

This document requests that a new sub-registry, named "STATEFUL-PCE-CAPABILITY TLV Flag Field", is created within the "Path Computation Element Protocol (PCEP) Numbers" registry to manage the Flag field in the STATEFUL-PCE-CAPABILITY TLV of the PCEP OPEN object (class = 1). New values are to be assigned by Standards Action [RFC5226]. Each bit should be tracked with the following qualities:

- o Bit number (counting from bit 0 as the most significant bit)
- o Capability description
- o Defining RFC

The following values are defined in this document:

Bit	Description	Reference
31	LSP-UPDATE-CAPABILITY	This document

### 8.9. LSP-ERROR-CODE TLV

This document requests that a new sub-registry, named "LSP-ERROR-CODE TLV Error Code Field", is created within the "Path Computation Element Protocol (PCEP) Numbers" registry to manage the LSP Error code field of the LSP-ERROR-CODE TLV. This field specifies the reason for failure to update the LSP.

New values are to be assigned by Standards Action [RFC5226]. Each value should be tracked with the following qualities: value, description and defining RFC. The following values are defined in this document:

Value	Meaning
1	Unknown reason
2	Limit reached for PCE-controlled LSPs
3	Too many pending LSP update requests
4	Unacceptable parameters
5	Internal error
6	LSP administratively brought down
7	LSP preempted
8	RSVP signaling error

## 9. Manageability Considerations

All manageability requirements and considerations listed in [RFC5440] apply to PCEP extensions defined in this document. In addition, requirements and considerations listed in this section apply.

### 9.1. Control Function and Policy

In addition to configuring specific PCEP session parameters, as specified in [RFC5440], Section 8.1, a PCE or PCC implementation MUST allow configuring the stateful PCEP capability and the LSP Update capability. A PCC implementation SHOULD allow the operator to specify multiple candidate PCEs for and a delegation preference for each candidate PCE. A PCC SHOULD allow the operator to specify an LSP delegation policy where LSPs are delegated to the most-preferred online PCE. A PCC MAY allow the operator to specify different LSP delegation policies.

A PCC implementation which allows concurrent connections to multiple PCEs SHOULD allow the operator to group the PCEs by administrative domains and it MUST NOT advertise LSP existence and state to a PCE if the LSP is delegated to a PCE in a different group.

A PCC implementation SHOULD allow the operator to specify whether the PCC will advertise LSP existence and state for LSPs that are not



controlled by any PCE (for example, LSPs that are statically configured at the PCC).

A PCC implementation SHOULD allow the operator to specify both the Redelegating Timeout Interval and the State Timeout Interval. The default value of the Redelegating Timeout Interval SHOULD be set to 30 seconds. An operator MAY also configure a policy that will dynamically adjust the Redelegating Timeout Interval, for example setting it to zero when the PCC has an established session to a backup PCE. The default value for the State Timeout Interval SHOULD be set to 60 seconds.

After the expiration of the State Timeout Interval, the LSP reverts to operator-defined default parameters. A PCC implementation MUST allow the operator to specify the default LSP parameters. To achieve a behavior where the LSP retains the parameters set by the PCE until such time that the PCC makes a change to them, a State Timeout Interval of infinity SHOULD be used. Any changes to LSP parameters SHOULD be done in make-before-break fashion.

LSP Delegation is controlled by operator-defined policies on a PCC. LSPs are delegated individually - different LSPs may be delegated to different PCEs. An LSP is delegated to at most one PCE at any given point in time. A PCC implementation SHOULD support the delegation policy, when all PCC's LSPs are delegated to a single PCE at any given time. Conversely, the policy revoking the delegation for all PCC's LSPs SHOULD also be supported.

A PCC implementation SHOULD allow the operator to specify delegation priority for PCEs. This effectively defines the primary PCE and one or more backup PCEs to which primary PCE's LSPs can be delegated when the primary PCE fails.

Policies defined for stateful PCEs and PCCs should eventually fit in the Policy-Enabled Path Computation Framework defined in [RFC5394], and the framework should be extended to support Stateful PCEs.

## 9.2. Information and Data Models

The PCEP YANG module [I-D.ietf-pcep-pcep-yang] should include

- o advertised stateful capabilities and synchronization status per PCEP session
- o the delegation status of each configured LSP.

The PCEP MIB [RFC7420] could also be updated to include this information.

### 9.3. Liveness Detection and Monitoring

PCEP extensions defined in this document do not require any new mechanisms beyond those already defined in [RFC5440], Section 8.3.

### 9.4. Verifying Correct Operation

Mechanisms defined in [RFC5440], Section 8.4 also apply to PCEP extensions defined in this document. In addition to monitoring parameters defined in [RFC5440], a stateful PCC-side PCEP implementation SHOULD provide the following parameters:

- o Total number of LSP updates
- o Number of successful LSP updates
- o Number of dropped LSP updates
- o Number of LSP updates where LSP setup failed

A PCC implementation SHOULD provide a command to show for each LSP whether it is delegated, and if so, to which PCE.

A PCC implementation SHOULD allow the operator to manually revoke LSP delegation.

### 9.5. Requirements on Other Protocols and Functional Components

PCEP extensions defined in this document do not put new requirements on other protocols.

### 9.6. Impact on Network Operation

Mechanisms defined in [RFC5440], Section 8.6 also apply to PCEP extensions defined in this document.

Additionally, a PCEP implementation SHOULD allow a limit to be placed on the number of LSPs delegated to the PCE and on the rate of PCUpd and PCRpt messages sent by a PCEP speaker and processed from a peer. It SHOULD also allow sending a notification when a rate threshold is reached.

A PCC implementation SHOULD allow a limit to be placed on the rate of LSP Updates to the same LSP to avoid signaling overload discussed in Section 10.3.

## 10. Security Considerations

### 10.1. Vulnerability

This document defines extensions to PCEP to enable stateful PCEs. The nature of these extensions and the delegation of path control to PCEs results in more information being available for a hypothetical adversary and a number of additional attack surfaces which must be protected.

The security provisions described in [RFC5440] remain applicable to these extensions. However, because the protocol modifications outlined in this document allow the PCE to control path computation timing and sequence, the PCE defense mechanisms described in [RFC5440] section 7.2 are also now applicable to PCC security.

As a general precaution, it is RECOMMENDED that these PCEP extensions only be activated on authenticated and encrypted sessions across PCEs and PCCs belonging to the same administrative authority, using Transport Layer Security (TLS) [I-D.ietf-pce-pceps], as per the recommendations and best current practices in [RFC7525].

The following sections identify specific security concerns that may result from the PCEP extensions outlined in this document along with recommended mechanisms to protect PCEP infrastructure against related attacks.

### 10.2. LSP State Snooping

The stateful nature of this extension explicitly requires LSP status updates to be sent from PCC to PCE. While this gives the PCE the ability to provide more optimal computations to the PCC, it also provides an adversary with the opportunity to eavesdrop on decisions made by network systems external to PCE. This is especially true if the PCC delegates LSPs to multiple PCEs simultaneously.

Adversaries may gain access to this information by eavesdropping on unsecured PCEP sessions, and might then use this information in various ways to target or optimize attacks on network infrastructure. For example by flexibly countering anti-DDoS measures being taken to protect the network, or by determining choke points in the network where the greatest harm might be caused.

PCC implementations which allow concurrent connections to multiple PCEs SHOULD allow the operator to group the PCEs by administrative domains and they MUST NOT advertise LSP existence and state to a PCE if the LSP is delegated to a PCE in a different group.

### 10.3. Malicious PCE

The LSP delegation mechanism described in this document allows a PCC to grant effective control of an LSP to the PCE for the duration of a PCEP session. While this enables PCE control of the timing and sequence of path computations within and across PCEP sessions, it also introduces a new attack vector: an attacker may flood the PCC with PCUpd messages at a rate which exceeds either the PCC's ability to process them or the network's ability to signal the changes, either by spoofing messages or by compromising the PCE itself.

A PCC is free to revoke an LSP delegation at any time without needing any justification. A defending PCC can do this by enqueueing the appropriate PCRpt message. As soon as that message is enqueued in the session, the PCC is free to drop any incoming PCUpd messages without additional processing.

### 10.4. Malicious PCC

A stateful session also results in an increased attack surface by placing a requirement for the PCE to keep an LSP state replica for each PCC. It is RECOMMENDED that PCE implementations provide a limit on resources a single PCC can occupy. A PCE implementing such a limit MUST send a PCNtf message with notification-type 4 (Stateful PCE resource limit exceeded) and notification-value 1 (Entering resource limit exceeded state) upon receiving an LSP state report causing it to exceed this threshold.

Delegation of LSPs can create further strain on PCE resources and a PCE implementation MAY preemptively give back delegations if it finds itself lacking the resources needed to effectively manage the delegation. Since the delegation state is ultimately controlled by the PCC, PCE implementations SHOULD provide throttling mechanisms to prevent strain created by flaps of either a PCEP session or an LSP delegation.

## 11. Contributing Authors

Xian Zhang  
Huawei Technology  
F3-5-B R&D Center  
Huawei Industrial Base, Bantian, Longgang District  
Shenzhen, Guangdong 518129  
P.R.China  
EMail: zhang.xian@huawei.com

Dhruv Dhody  
Huawei Technology

Leela Palace  
Bangalore, Karnataka 560008  
INDIA  
EMail: dhruv.dhody@huawei.com

Siva Sivabalan  
Cisco Systems, Inc.  
2000 Innovation Drive  
Kanata, Ontario K2K 3E8  
Canada  
EMail: msiva@cisco.com

## 12. Acknowledgements

We would like to thank Adrian Farrel, Cyril Margaria and Ramon Casellas for their contributions to this document.

We would like to thank Shane Amante, Julien Meuric, Kohei Shiimoto, Paul Schultz and Raveendra Torvi for their comments and suggestions. Thanks also to Jon Hardwick, Oscar Gonzales de Dios, Tomas Janciga, Stefan Kobza, Kexin Tang, Matej Spanik, Jon Parker, Marek Zavodsky, Ambrose Kwong, Ashwin Sampath, Calvin Ying, Mustapha Aissaoui, Stephane Litkowski and Olivier Dugeon for helpful comments and discussions.

## 13. References

### 13.1. Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <<http://www.rfc-editor.org/info/rfc2119>>.
- [RFC2205] Braden, R., Ed., Zhang, L., Berson, S., Herzog, S., and S. Jamin, "Resource ReSerVation Protocol (RSVP) -- Version 1 Functional Specification", RFC 2205, DOI 10.17487/RFC2205, September 1997, <<http://www.rfc-editor.org/info/rfc2205>>.
- [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, <<http://www.rfc-editor.org/info/rfc3209>>.
- [RFC5088] Le Roux, JL., Ed., Vasseur, JP., Ed., Ikejiri, Y., and R. Zhang, "OSPF Protocol Extensions for Path Computation Element (PCE) Discovery", RFC 5088, DOI 10.17487/RFC5088, January 2008, <<http://www.rfc-editor.org/info/rfc5088>>.

- [RFC5089] Le Roux, JL., Ed., Vasseur, JP., Ed., Ikejiri, Y., and R. Zhang, "IS-IS Protocol Extensions for Path Computation Element (PCE) Discovery", RFC 5089, DOI 10.17487/RFC5089, January 2008, <<http://www.rfc-editor.org/info/rfc5089>>.
- [RFC5284] Swallow, G. and A. Farrel, "User-Defined Errors for RSVP", RFC 5284, DOI 10.17487/RFC5284, August 2008, <<http://www.rfc-editor.org/info/rfc5284>>.
- [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>>.
- [RFC5511] Farrel, A., "Routing Backus-Naur Form (RBNF): A Syntax Used to Form Encoding Rules in Various Routing Protocol Specifications", RFC 5511, DOI 10.17487/RFC5511, April 2009, <<http://www.rfc-editor.org/info/rfc5511>>.
- [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>>.

### 13.2. Informative References

- [I-D.ietf-pce-gmpls-pcep-extensions]  
Margarita, C., Dios, O., and F. Zhang, "PCEP extensions for GMPLS", draft-ietf-pce-gmpls-pcep-extensions-11 (work in progress), October 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-pcep-yang]  
Dhody, D., Hardwick, J., Beeram, V., and j. jeffrant@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-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-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.
- [MPLS-PC] Chaieb, I., Le Roux, J.L., and B. Cousin, "Improved MPLS-TE  
LSP Path Computation using Preemption", Global  
Information Infrastructure Symposium, July 2007.
- [MXMN-TE] Danna, E., Mandal, S., and A. Singh, "Practical linear  
programming algorithm for balancing the max-min fairness  
and throughput objectives in traffic engineering",  
INFOCOM, 2012 Proceedings IEEE Page(s): 846-854, 2012.
- [RFC2702] Awduche, D., Malcolm, J., Agogbua, J., O'Dell, M., and J.  
McManus, "Requirements for Traffic Engineering Over MPLS",  
RFC 2702, DOI 10.17487/RFC2702, September 1999,  
<<http://www.rfc-editor.org/info/rfc2702>>.
- [RFC3031] Rosen, E., Viswanathan, A., and R. Callon, "Multiprotocol  
Label Switching Architecture", RFC 3031,  
DOI 10.17487/RFC3031, January 2001,  
<<http://www.rfc-editor.org/info/rfc3031>>.
- [RFC3346] Boyle, J., Gill, V., Hannan, A., Cooper, D., Awduche, D.,  
Christian, B., and W. Lai, "Applicability Statement for  
Traffic Engineering with MPLS", RFC 3346,  
DOI 10.17487/RFC3346, August 2002,  
<<http://www.rfc-editor.org/info/rfc3346>>.
- [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>>.
- [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>>.
- [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, <<http://www.rfc-editor.org/info/rfc4657>>.

- [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>>.
- [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>>.
- [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>>.
- [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>>.
- [RFC7525] Sheffer, Y., Holz, R., and P. Saint-Andre, "Recommendations for Secure Use of Transport Layer Security (TLS) and Datagram Transport Layer Security (DTLS)", BCP 195, RFC 7525, DOI 10.17487/RFC7525, May 2015, <<http://www.rfc-editor.org/info/rfc7525>>.

#### Authors' Addresses

Edward Crabbe  
Oracle  
1501 4th Ave, suite 1800  
Seattle, WA 98101  
US  
  
Email: [edward.crabbe@oracle.com](mailto:edward.crabbe@oracle.com)

Ina Minei  
Google, Inc.  
1600 Amphitheatre Parkway  
Mountain View, CA 94043  
US  
  
Email: [inaminei@google.com](mailto:inaminei@google.com)



Jan Medved  
Cisco Systems, Inc.  
170 West Tasman Dr.  
San Jose, CA 95134  
US

Email: [jmedved@cisco.com](mailto:jmedved@cisco.com)

Robert Varga  
Pantheon Technologies SRO  
Mlynske Nivy 56  
Bratislava 821 05  
Slovakia

Email: [robert.varga@pantheon.tech](mailto:robert.varga@pantheon.tech)

Network Working Group  
Internet Draft  
Intended status: Informational  
Expires: April 2015

Y. Lee  
Huawei  
G. Bernstein  
Grotto Networking  
Jonas Martensson  
Acreo  
T. Takeda  
NTT  
T. Tsuritani  
KDDI  
O. G. de Dios  
Telefonica

October 28, 2014

## PCEP Requirements for WSON Routing and Wavelength Assignment

draft-ietf-pce-wson-routing-wavelength-15.txt

### Abstract

This memo provides application-specific requirements for the Path Computation Element communication Protocol (PCEP) for the support of Wavelength Switched Optical Networks (WSON). Lightpath provisioning in WSONs requires a routing and wavelength assignment (RWA) process. From a path computation perspective, wavelength assignment is the process of determining which wavelength can be used on each hop of a path and forms an additional routing constraint to optical light path computation. Requirements for PCEP extensions in support of optical impairments will be addressed in a separate document.

### Status of this Memo

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

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

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

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

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

This Internet-Draft will expire on April 28, 2009.

#### Copyright Notice

Copyright (c) 2014 IETF Trust and the persons identified as the document authors. All rights reserved.

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

#### Conventions used in this document

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

#### Table of Contents

1. Introduction.....	3
2. WSON RWA Processes & Architecture.....	4
3. Requirements.....	6
3.1. Path Computation Type Option.....	6
3.2. RWA Processing.....	6
3.3. Bulk RWA Path Request/Reply.....	7
3.4. RWA Path Re-optimization Request/Reply.....	7
3.5. Wavelength Range Constraint.....	8
3.6. Wavelength Assignment Preference.....	8
3.7. Signal Processing Capability Restriction.....	8
4. Manageability Considerations.....	9
4.1. Control of Function and Policy.....	9
4.2. Information and Data Models, e.g. MIB module.....	9
4.3. Liveness Detection and Monitoring.....	10
4.4. Verifying Correct Operation.....	10

4.5. Requirements on Other Protocols and Functional Components	10
4.6. Impact on Network Operation	10
5. Security Considerations	10
6. IANA Considerations	11
7. Acknowledgments	11
8. References	11
8.1. Normative References	11
8.2. Informative References	11
Authors' Addresses	12
Intellectual Property Statement	13
Disclaimer of Validity	13

## 1. Introduction

[RFC4655] defines the PCE-based architecture and explains how a Path Computation Element (PCE) may compute Label Switched Paths (LSP) in Multiprotocol Label Switching Traffic Engineering (MPLS-TE) and Generalized MPLS (GMPLS)-controlled networks at the request of Path Computation Clients (PCCs). A PCC is shown to be any network component that makes such a request and may be for instance an optical switching element within a Wavelength Division Multiplexing (WDM) network. The PCE, itself, can be located anywhere within the network, and may be within an optical switching element, a Network Management System (NMS) or Operational Support System (OSS), or may be an independent network server.

The PCE communication Protocol (PCEP) is the communication protocol used between PCC and PCE, and may also be used between cooperating PCEs. [RFC4657] sets out the common protocol requirements for PCEP. Additional application-specific requirements for PCEP are deferred to separate documents.

This document provides a set of application-specific PCEP requirements for support of path computation in Wavelength Switched Optical Networks (WSON). WSON refers to WDM-based optical networks in which switching is performed selectively based on the wavelength of an optical signal.

The path in WSON is referred to as a lightpath. A lightpath may span multiple fiber links and the path should be assigned a wavelength for each link.

A transparent optical network is made up of optical devices that can switch but not convert from one wavelength to another. In a transparent optical network, a lightpath operates on the same wavelength across all fiber links that it traverses. In such case, the lightpath is said to satisfy the wavelength-continuity

constraint. Two lightpaths that share a common fiber link cannot be assigned the same wavelength. To do otherwise would result in both signals interfering with each other. Note that advanced additional multiplexing techniques such as polarization based multiplexing are not addressed in this document since the physical layer aspects are not currently standardized. Therefore, assigning the proper wavelength on a lightpath is an essential requirement in the optical path computation process.

When a switching node has the ability to perform wavelength conversion the wavelength-continuity constraint can be relaxed, and a lightpath may use different wavelengths on different links along its path from origin to destination. It is, however, to be noted that wavelength converters may be limited for cost reasons, while the number of WDM channels that can be supported in a fiber is also limited. As a WSON can be composed of network nodes that cannot perform wavelength conversion, nodes with limited wavelength conversion, and nodes with full wavelength conversion abilities, wavelength assignment is an additional routing constraint to be considered in all lightpath computations.

In this document we first review the processes for routing and wavelength assignment (RWA) used when wavelength continuity constraints are present and then specify requirements for PCEP to support RWA. Requirements for optical impairments will be addressed in a separate document.

The remainder of this document uses terminology from [RFC4655].

## 2. WSON RWA Processes & Architecture

In [RFC6163] three alternative process architectures were given for performing routing and wavelength assignment. These are shown schematically in Figure 1. R stands for Routing, WA for Wavelength Assignment, and DWA for Distributed Wavelength Assignment.

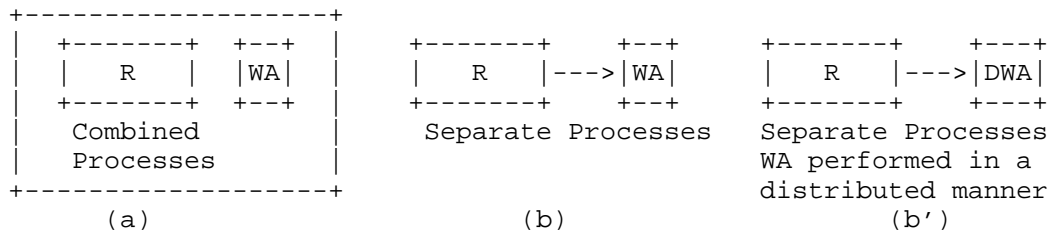


Figure 1. RWA process alternatives

These alternatives have the following properties and impact on PCEP requirements in this document.

(a) Combined Processes (R&WA)

Here path selection and wavelength assignment are performed as a single process. The requirements for PCC-PCE interaction with such a combined RWA process PCE is addressed in this document.

(b) Routing separate from Wavelength Assignment (R+WA)

Here the routing process furnishes one or more potential paths to the wavelength assignment process that then performs final path selection and wavelength assignment. The requirements for PCE-PCE interaction with one PCE implementing the routing process and another implementing the wavelength assignment process are not addressed in this document.

(b') Routing and distributed Wavelength Assignment (R+DWA)

Here a standard path computation (unaware of detailed wavelength availability) takes place, then wavelength assignment is performed along this path in a distributed manner via signaling (RSVP-TE). This alternative is a particular case of R+WA and it should be covered by GMPLS PCEP extensions and does not present new WSON-specific requirements.

In the previous section various process architectures for implementing RWA have been reviewed. Figure 2 shows one typical PCE-based implementation, which is referred to as Combined Process (R&WA). With this architecture, the two processes of routing and wavelength assignment are accessed via a single PCE. This architecture is the base architecture from which the requirements are specified in this document.

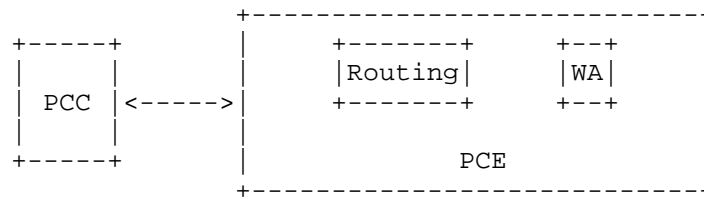


Figure 2. Combined Process (R&amp;WA) architecture

### 3. Requirements

The requirements for the PCC to PCE interface of Figure 2 are specified in this section.

#### 3.1. Path Computation Type Option

A PCEP request MAY include the path computation type. This can be:

- (i) Both Routing and Wavelength Assignment (RWA),
- (ii) Routing only.

This requirement is needed to differentiate between the currently supported routing with distributed wavelength assignment option and combined RWA. In case of distributed wavelength assignment option, wavelength assignment will be performed at each node of the route.

#### 3.2. RWA Processing

- (a) When the request is a RWA path computation type, the request MUST further include the wavelength assignment options. At the minimum, the following option should be supported:

- (i) Explicit Label Control (ELC) [RFC3473]
- (ii) A set of recommended labels for each hop. The PCC can select the label based on local policy.

Note that option (ii) may also be used in R+WA or R+DWA.

- (b) In case of a RWA computation type, the response MUST include the wavelength(s) assigned to the path and an indication of which label assignment option has been applied (ELC or label set).

- (c) In the case where a valid path is not found, the response MUST include why the path is not found (e.g., network disconnected, wavelength not found, or both, etc.). Note that 'wavelength not found' may include several sub-cases such as wavelength continuity not met, unsupported FEC/Modulation type, etc.

### 3.3. Bulk RWA Path Request/Reply

Sending simultaneous path requests for "routing only" computation is supported by PCEP specification [RFC5440]. To remain consistent the following requirements are added.

- (a) A PCEP request MUST be able to specify an option for bulk RWA path request. Bulk path request is an ability to request a number of simultaneous RWA path requests.
- (b) The PCEP response MUST include the path and the assigned wavelength assigned for each RWA path request specified in the original bulk request.

### 3.4. RWA Path Re-optimization Request/Reply

1. For a re-optimization request, the request MUST provide both the path and current wavelength to be re-optimized and MAY include the following options:
  - a. Re-optimize the path keeping the same wavelength(s)
  - b. Re-optimize wavelength(s) keeping the same path
  - c. Re-optimize allowing both the wavelength and the path to change
2. The corresponding response to the re-optimized request MUST provide the re-optimized path and wavelengths even when the request asked for the path or the wavelength to remain unchanged.
3. In case that the new path is not found, the response MUST include why the path is not found (e.g., network disconnected, wavelength not found, or both, etc.). Note that 'wavelength not found' may include several sub-cases such as wavelength continuity not met, unsupported FEC/Modulation type, etc.



### 3.5. Wavelength Range Constraint

For any RWA computation type request, the requester (PCC) MUST be allowed to specify a restriction on the wavelengths to be used. The requester MAY use this option to restrict the assigned wavelength for explicit label or label set. This restriction may for example come from the tuning ability of a laser transmitter, any optical element, or a policy-based restriction.

Note that the requester (e.g., PCC) is not required to furnish any range restrictions.

### 3.6. Wavelength Assignment Preference

1. A RWA computation type request MAY include the requester preference for, e.g., random assignment, descending order, ascending order, etc. A response SHOULD follow the requestor preference unless it conflicts with operator's policy.
2. A request for two or more paths MUST allow the requester to include an option constraining the paths to have the same wavelength(s) assigned. This is useful in the case of protection with single transponder (e.g., 1+1 link disjoint paths).

In a network with wavelength conversion capabilities (e.g. sparse 3R regenerators), a request SHOULD be able to indicate whether a single, continuous wavelength should be allocated or not. In other words, the requesting PCC SHOULD be able to specify the precedence of wavelength continuity even if wavelength conversion is available.

### 3.7. Signal Processing Capability Restriction

Signal processing compatibility is an important constraint for optical path computation. The signal type for an end-to-end optical path must match at source and at destination.

The PCC MUST be allowed to specify the signal type at the endpoints (i.e., at source and at destination). The following signal processing capabilities should be supported at a minimum:

- o Modulation Type List
- o FEC Type List

The PCC MUST also be allowed to state whether transit modification is acceptable for the above signal processing capabilities.

#### 4. Manageability Considerations

Manageability of WSON Routing and Wavelength Assignment (RWA) with PCE must address the following considerations:

##### 4.1. Control of Function and Policy

In addition to the parameters already listed in Section 8.1 of [RFC5440], a PCEP implementation SHOULD allow configuring the following PCEP session parameters on a PCC:

- o The ability to send a WSON RWA request.

In addition to the parameters already listed in Section 8.1 of [RFC5440], a PCEP implementation SHOULD allow configuring the following PCEP session parameters on a PCE:

- o The support for WSON RWA.
- o The maximum number of bulk path requests associated with WSON RWA per request message.

These parameters may be configured as default parameters for any PCEP session the PCEP speaker participates in, or may apply to a specific session with a given PCEP peer or a specific group of sessions with a specific group of PCEP peers.

##### 4.2. Information and Data Models, e.g. MIB module

As this document only concerns the requirements to support WSON RWA, no additional MIB module is defined in this document. However, the corresponding solution draft will list the information that should be added to the PCE MIB module defined in [PCEP-MIB].

#### 4.3. Liveness Detection and Monitoring

No new mechanism is defined in this document that implies any new liveness detection and monitoring requirements in addition to those already listed in section 8.3 of [RFC5440].

#### 4.4. Verifying Correct Operation

No new mechanism is defined in this document that implies any new verification requirements in addition to those already listed in section 8.4 of [RFC5440]

#### 4.5. Requirements on Other Protocols and Functional Components

If PCE discovery mechanisms ([RFC5089] and [RFC5088]) were to be extended for technology-specific capabilities, advertising WSON RWA path computation capability should be considered.

#### 4.6. Impact on Network Operation

No new mechanism is defined in this document that implies any new network operation requirements in addition to those already listed in section 8.6 of [RFC5440].

### 5. Security Considerations

This document has no requirement for a change to the security models within PCEP [RFC5440]. However the additional information distributed in order to address the RWA problem represents a disclosure of network capabilities that an operator may wish to keep private. Consideration should be given to securing this information.

Solutions that address the requirements in this document need to verify that existing PCEP security mechanisms adequately protect the additional network capabilities and must include new mechanisms as necessary.

## 6. IANA Considerations

This informational document does not make any requests for IANA action.

## 7. Acknowledgments

The authors would like to thank Adrian Farrel, Cycil Margaria and Ramon Casellas for many helpful comments that greatly improved the contents of this draft.

This document was prepared using 2-Word-v2.0.template.dot.

## 8. References

### 8.1. Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC4655] Farrel, A., Vasseur, J., and J. Ash, "A Path Computation Element (PCE)-Based Architecture", RFC 4655, August 2006.
- [RFC5440] Vasseur, JP., Ed. and JL. Le Roux, Ed., "Path Computation Element (PCE) communication Protocol", RFC 5440, March 2009.

### 8.2. Informative References

- [RFC3473] L. Berger, "Generalized Multi-Protocol Label Switching (GMPLS) Signaling Resource ReserVation Protocol-Traffic Engineering (RSVP-TE) Extensions", RFC 3473, January 2003.
- [RFC4657] Ash, J. and J. Le Roux, "Path Computation Element (PCE) Communication Protocol Generic Requirements", RFC 4657, September 2006.

- [RFC6163] Y. Lee, G. Bernstein, W. Imajuku, "Framework for GMPLS and PCE Control of Wavelength Switched Optical Networks", RFC 6163, April 2011.
- [RFC5088] Le Roux, JL., Ed., Vasseur, JP., Ed., Ikejiri, Y., and R. Zhang, "OSPF Protocol Extensions for Path Computation Element (PCE) Discovery", RFC 5088, January 2008.
- [RFC5089] Le Roux, JL., Ed., Vasseur, JP., Ed., Ikejiri, Y., and R. Zhang, "IS-IS Protocol Extensions for Path Computation Element (PCE) Discovery", RFC 5089, January 2008.
- [PCEP-MIB] Koushik, K, et al., "PCE communication protocol(PCEP) Management Information Base", draft-ietf-pce-pcep-mib, work in progress.

#### Authors' Addresses

Young Lee (Ed.)  
Huawei Technologies  
5340 Legacy Drive, Building 3  
Plano, TX 75245, USA  
Phone: (469)277-5838  
Email: leeyoung@huawei.com

Greg Bernstein (Ed.)  
Grotto Networking  
Fremont, CA, USA  
Phone: (510) 573-2237  
Email: gregb@grotto-networking.com

Jonas Martensson  
Acreo  
Email:Jonas.Martensson@acreo.se

Tomonori Takeda  
NTT Corporation  
3-9-11, Midori-Cho  
Musashino-Shi, Tokyo 180-8585, Japan  
Email: takeda.tomonori@lab.ntt.co.jp

Takehiro Tsuritani  
KDDI R&D Laboratories, Inc.  
2-1-15 Ohara Kamifukuoka Saitama, 356-8502. Japan  
Phone: +81-49-278-7357  
Email: tsuri@kddilabs.jp

Oscar Gonzalez de Dios  
Telefonica Investigacion y Desarrollo  
C/ Emilio Vargas 6  
Madrid, 28043  
Spain  
Phone: +34 91 3374013  
Email: ogondio@tid.es



Network Working Group  
Internet Draft

Y. Lee, Ed.  
Huawei Technologies

Intended status: Standard Track  
Expires: September 1, 2019

R. Casellas, Ed.  
CTTC

March 1, 2019

## PCEP Extension for WSON Routing and Wavelength Assignment

draft-ietf-pce-wson-rwa-ext-17

### Abstract

This document provides the Path Computation Element communication Protocol (PCEP) extensions for the support of Routing and Wavelength Assignment (RWA) in Wavelength Switched Optical Networks (WSON). Path provisioning in WSONs requires a routing and wavelength assignment (RWA) process. From a path computation perspective, wavelength assignment is the process of determining which wavelength can be used on each hop of a path and forms an additional routing constraint to optical path computation.

### Status of this Memo

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

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

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

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



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

This Internet-Draft will expire on September 1, 2019.

#### Copyright Notice

Copyright (c) 2019 IETF Trust and the persons identified as the document authors. All rights reserved.

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

#### Table of Contents

1. Terminology.....	3
2. Requirements Language.....	3
3. Introduction.....	3
4. Encoding of a RWA Path Request.....	6
4.1. Wavelength Assignment (WA) Object.....	7
4.2. Wavelength Selection TLV.....	9
4.3. Wavelength Restriction Constraint TLV.....	9
4.3.1. Link Identifier Field.....	12
4.3.2. Wavelength Restriction Field.....	14
4.4. Signal Processing Capability Restrictions.....	15
4.4.1. Signal Processing Exclusion.....	16
4.4.2. Signal Processing Inclusion.....	18
5. Encoding of a RWA Path Reply.....	19
5.1. Wavelength Allocation TLV.....	19
5.2. Error Indicator.....	20
5.3. NO-PATH Indicator.....	21
6. Manageability Considerations.....	22
6.1. Control of Function and Policy.....	22
6.2. Liveness Detection and Monitoring.....	22
6.3. Verifying Correct Operation.....	22
6.4. Requirements on Other Protocols and Functional Components.....	22
6.5. Impact on Network Operation.....	23

7. Security Considerations.....	23
8. IANA Considerations.....	23
8.1. New PCEP Object: Wavelength Assignment Object.....	23
8.2. WA Object Flag Field.....	23
8.3. New PCEP TLV: Wavelength Selection TLV.....	24
8.4. New PCEP TLV: Wavelength Restriction Constraint TLV.....	24
8.5. Wavelength Restriction Constraint TLV Action Values.....	25
8.6. New PCEP TLV: Wavelength Allocation TLV.....	25
8.7. Wavelength Allocation TLV Flag Field.....	25
8.8. New PCEP TLV: Optical Interface Class List TLV.....	26
8.9. New PCEP TLV: Client Signal TLV.....	26
8.10. New No-Path Reasons.....	27
8.11. New Error-Types and Error-Values.....	27
8.12. New Subobjects for the Exclude Route Object.....	28
8.13. New Subobjects for the Include Route Object.....	28
8.14. Request for Updated Note for LMP TE Link Object Class Type .....	28
9. Acknowledgments.....	29
10. References.....	29
10.1. Normative References.....	29
10.2. Informative References.....	30
11. Contributors.....	32
Authors' Addresses.....	33

## 1. Terminology

This document uses the terminology defined in [RFC4655], and [RFC5440].

## 2. Requirements Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "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.

## 3. Introduction

[RFC5440] specifies the Path Computation Element (PCE) Communication Protocol (PCEP) for communications between a Path Computation Client (PCC) and a PCE, or between two PCEs. Such interactions include path computation requests and path computation replies as well as notifications of specific states related to the use of a PCE in the

context of Multiprotocol Label Switching (MPLS) and Generalized MPLS (GMPLS) Traffic Engineering.

A PCC is said to be any network component that makes such a request and may be, for instance, an Optical Switching Element within a Wavelength Division Multiplexing (WDM) network. The PCE, itself, can be located anywhere within the network, and may be within an optical switching element, a Network Management System (NMS) or Operational Support System (OSS), or may be an independent network server.

This document provides the PCEP extensions for the support of Routing and Wavelength Assignment (RWA) in Wavelength Switched Optical Networks (WSON) based on the requirements specified in [RFC6163] and [RFC7449].

WSON refers to WDM based optical networks in which switching is performed selectively based on the wavelength of an optical signal. The devices used in WSONs that are able to switch signals based on signal wavelength are known as Lambda Switch Capable (LSC). WSONs can be transparent or translucent. A transparent optical network is made up of optical devices that can switch but not convert from one wavelength to another, all within the optical domain. On the other hand, translucent networks include 3R regenerators (Re-amplification, Re-shaping, Re-timing) that are sparsely placed. The main function of the 3R regenerators is to convert one optical wavelength to another.

A Lambda Switch Capable (LSC) Label Switched Path (LSP) may span one or several transparent segments, which are delimited by 3R regenerators typically with electronic regenerator and optional wavelength conversion. Each transparent segment or path in WSON is referred to as an optical path. An optical path may span multiple fiber links and the path should be assigned the same wavelength for each link. In such case, the optical path is said to satisfy the wavelength-continuity constraint. Figure 1 illustrates the relationship between a LSC LSP and transparent segments (optical paths).

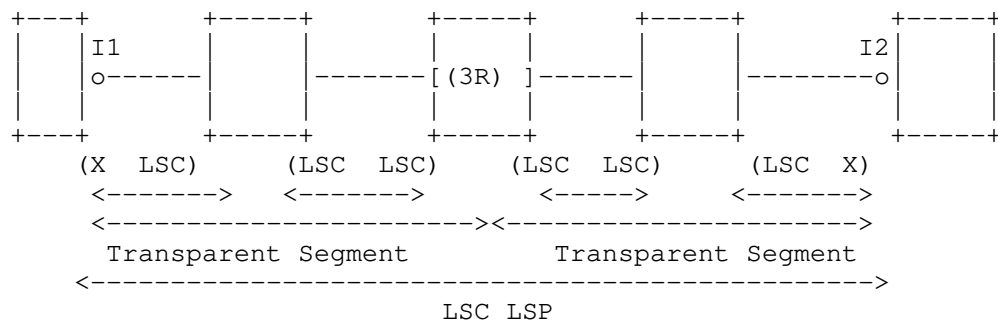


Figure 1 Illustration of a LSC LSP and transparent segments

Note that two transparent segments within a WSON LSP do not need to operate on the same wavelength (due to the wavelength conversion capabilities). Two optical channels that share a common fiber link cannot be assigned the same wavelength; Otherwise, the two signals would interfere with each other. Note that advanced additional multiplexing techniques such as polarization based multiplexing are not addressed in this document since the physical layer aspects are not currently standardized. Therefore, assigning the proper wavelength on a path is an essential requirement in the optical path computation process.

When a switching node has the ability to perform wavelength conversion, the wavelength-continuity constraint can be relaxed, and a LSC Label Switched Path (LSP) may use different wavelengths on different links along its route from origin to destination. It is, however, to be noted that wavelength converters may be limited due to their relatively high cost, while the number of WDM channels that can be supported in a fiber is also limited. As a WSON can be composed of network nodes that cannot perform wavelength conversion, nodes with limited wavelength conversion, and nodes with full wavelength conversion abilities, wavelength assignment is an additional routing constraint to be considered in all optical path computation.

For example (see Figure 1), within a translucent WSON, a LSC LSP may be established between interfaces I1 and I2, spanning 2 transparent segments (optical paths) where the wavelength continuity constraint applies (i.e. the same unique wavelength must be assigned to the LSP at each TE link of the segment). If the LSC LSP induced a Forwarding Adjacency / TE link, the switching capabilities of the TE link would

be (X X) where X refers to the switching capability of I1 and I2. For example, X can be Packet Switch Capable (PSC), Time Division Multiplexing (TDM), etc.

This document aligns with GMPLS extensions for PCEP [PCEP-GMPLS] for generic properties such as label, label-set and label assignment noting that wavelength is a type of label. Wavelength restrictions and constraints are also formulated in terms of labels per [RFC7579].

The optical modulation properties, which are also referred to as signal compatibility, are already considered in signaling in [RFC7581] and [RFC7688]. In order to improve the signal quality and limit some optical effects several advanced modulation processing capabilities are used by the mechanisms specified in this document. These modulation capabilities contribute not only to optical signal quality checks but also constrain the selection of sender and receiver, as they should have matching signal processing capabilities. This document includes signal compatibility constraints as part of RWA path computation. That is, the signal processing capabilities (e.g., modulation and Forward Error Correction (FEC)) indicated by means of optical interface class (OIC) must be compatible between the sender and the receiver of the optical path across all optical elements.

This document, however, does not address optical impairments as part of RWA path computation. See [RFC6566] for the framework for optical impairments.

#### 4. Encoding of a RWA Path Request

Figure 2 shows one typical PCE based implementation, which is referred to as the Combined Process (R&WA). With this architecture, the two processes of routing and wavelength assignment are accessed via a single PCE. This architecture is the base architecture specified in [RFC6163] and the PCEP extensions that are specified in this document are based on this architecture.

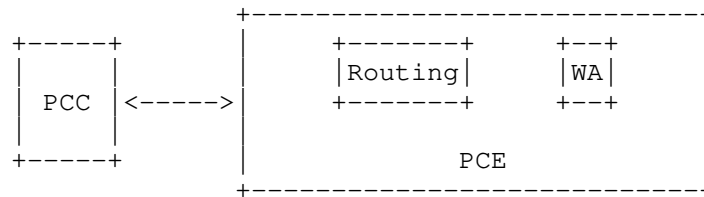


Figure 2 Combined Process (R&amp;WA) architecture

#### 4.1. Wavelength Assignment (WA) Object

Wavelength allocation can be performed by the PCE by different means:

(a) By means of Explicit Label Control [RFC3471] where the PCE allocates which label to use for each interface/node along the path. The allocated labels MAY appear after an interface route subobject.

(b) By means of a Label Set where the PCE provides a range of potential labels to allocate by each node along the path.

Option (b) allows distributed label allocation (performed during signaling) to complete wavelength assignment.

Additionally, given a range of potential labels to allocate, a PC Request SHOULD convey the heuristic / mechanism used for the allocation.

The format of a PCReq message per [RFC5440] after incorporating the Wavelength Assignment (WA) object is as follows:

```

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

```

Where:

```

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

```

<END-POINTS>

<WA>

[other optional objects...]

If the WA object is present in the request, it MUST be encoded after the END-POINTS object as defined in [PCEP-GMPLS]. The WA Object is mandatory in this document. Orderings for the other optional objects are irrelevant.

WA Object-Class is (TBD1) (To be assigned by IANA).

WA Object-Type is 1.

The format of the WA object body is as follows:

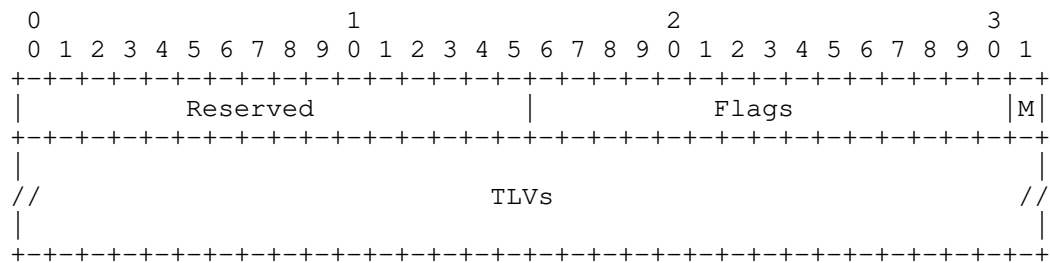


Figure 3 WA Object

- o Reserved (16 bits): Reserved for future use and SHOULD be zeroed and ignored on receipt.
- o Flags (16 bits)

One flag bit is allocated as follows:

- M (Mode - 1 bit): M bit is used to indicate the mode of wavelength assignment. When M bit is set to 1, this indicates that the label assigned by the PCE must be explicit. That is, the selected way to convey the allocated wavelength is by means of Explicit Label Control for each hop of a computed LSP. Otherwise (M bit is set to 0), the label assigned by the PCE need not be explicit (i.e., it can be suggested in the form of label set objects in the corresponding response, to allow distributed WA. If M is 0, the PCE MUST return a Label Set Field as described in Section 2.6 of [RFC7579] in the response. See Section 5 of this document for the encoding discussion of a Label Set Field in a PCRep message.

All unused flags SHOULD be zeroed. IANA is to create a new registry to manage the Flag field of the WA object.

- o TLVs (variable). In the TLVs field, the following two TLVs are defined. At least one TLV MUST be present.
  - Wavelength Selection TLV: A TLV of type (TBD2) with fixed length of 32 bits indicating the wavelength selection. See Section 4.2 for details.
  - Wavelength Restriction Constraint TLV: A TLV of type (TBD3) with variable length indicating wavelength restrictions. See Section 4.3 for details.

#### 4.2. Wavelength Selection TLV

The Wavelength Selection TLV is used to indicate the wavelength selection constraint in regard to the order of wavelength assignment to be returned by the PCE. This TLV is only applied when M bit is set in the WA Object specified in Section 4.1. This TLV MUST NOT be used when the M bit is cleared.

The encoding of this TLV is specified as the Wavelength Selection Sub-TLV in Section 4.2.2 of [RFC7689]. IANA is to allocate a new TLV type, Wavelength Selection TLV type (TBD2).

#### 4.3. Wavelength Restriction Constraint TLV

For any request that contains a wavelength assignment, the requester (PCC) MUST specify a restriction on the wavelengths to be used. This restriction is to be interpreted by the PCE as a constraint on the tuning ability of the origination laser transmitter or on any other



maintenance related constraints. Note that if the LSP LSC spans different segments, the PCE must have mechanisms to know the tunability restrictions of the involved wavelength converters / regenerators, e.g. by means of the Traffic Engineering Database (TED) either via IGP or Network Management System (NMS). Even if the PCE knows the tunability of the transmitter, the PCC must be able to apply additional constraints to the request.

The format of the Wavelength Restriction Constraint TLV is as follows:

```
<Wavelength Restriction Constraint> ::=
    (<Action> <Count> <Reserved>
     <Link Identifiers> <Wavelength Restriction>)...

```

Where

```
<Link Identifiers> ::= <Link Identifier> [<Link Identifiers>]

```

See Section 4.3.1. for the encoding of the Link Identifiers Field.

These fields (i.e., <Action>, <Link Identifiers> and <Wavelength Restriction>, etc.) MAY appear together more than once to be able to specify multiple actions and their restrictions.

IANA is to allocate a new TLV type, Wavelength Restriction Constraint TLV type (TBD3).

The TLV data is defined as follows:

```

0                               1                               2                               3
0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
| Action          |          Count          |          Reserved          |
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
|                                     Link Identifiers Field                                     |
//                                     . . .                                     //
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
|                                     Wavelength Restriction Field                                     |
//                                     . . . .                                     //

```

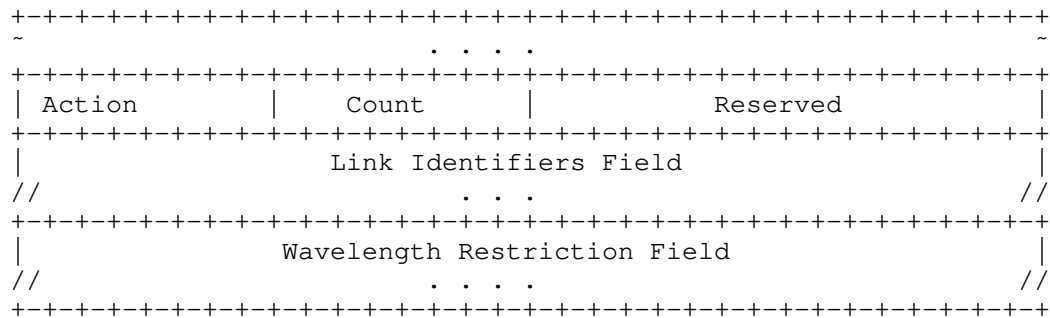


Figure 4 Wavelength Restriction Constraint TLV Encoding

- o Action (8 bits):

- o 0 - Inclusive List indicates that one or more link identifiers are included in the Link Set. Each identifies a separate link that is part of the set.
- o 1 - Inclusive Range indicates that the Link Set defines a range of links. It contains two link identifiers. The first identifier indicates the start of the range (inclusive). The second identifier indicates the end of the range (inclusive). All links with numeric values between the bounds are considered to be part of the set. A value of zero in either position indicates that there is no bound on the corresponding portion of the range.
- o 2-255 - For future use

IANA is to create a new registry to manage the Action values of the Wavelength Restriction Constraint TLV.

If PCE receives an unrecognized Action value, the PCE MUST send a PCERR message with a PCEP-ERROR Object (Error-Type=TBD8) and an Error-value (Error-value=3). See Section 5.2 for details.

Note that "links" are assumed to be bidirectional.

- o Count (8 bits): The number of the link identifiers

Note that a PCC MAY add a Wavelength restriction that applies to all links by setting the Count field to zero and specifying just a set of wavelengths.

Note that all link identifiers in the same list MUST be of the same type.

- o Reserved (16 bits): Reserved for future use and SHOULD be zeroed and ignored on receipt.
- o Link Identifiers: Identifies each link ID for which restriction is applied. The length is dependent on the link format and the Count field. See Section 4.3.1. for Link Identifier encoding.
- o Wavelength Restriction: See Section 4.3.2. for the Wavelength Restriction Field encoding.

Various encoding errors are possible with this TLV (e.g., not exactly two link identifiers with the range case, unknown identifier types, no matching link for a given identifier, etc.). To indicate errors associated with this encoding, a PCEP speaker MUST send a PCErr message with Error-Type=TBD8 and Error-value=3. See Section 5.1 for the details.

#### 4.3.1. Link Identifier Field

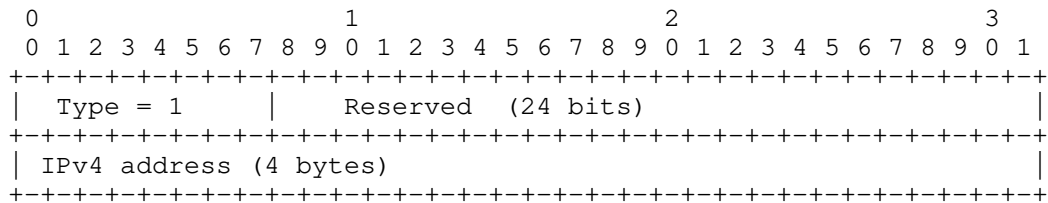
The link identifier field can be an IPv4 [RFC3630], IPv6 [RFC5329] or unnumbered interface ID [RFC4203].

<Link Identifier> ::=

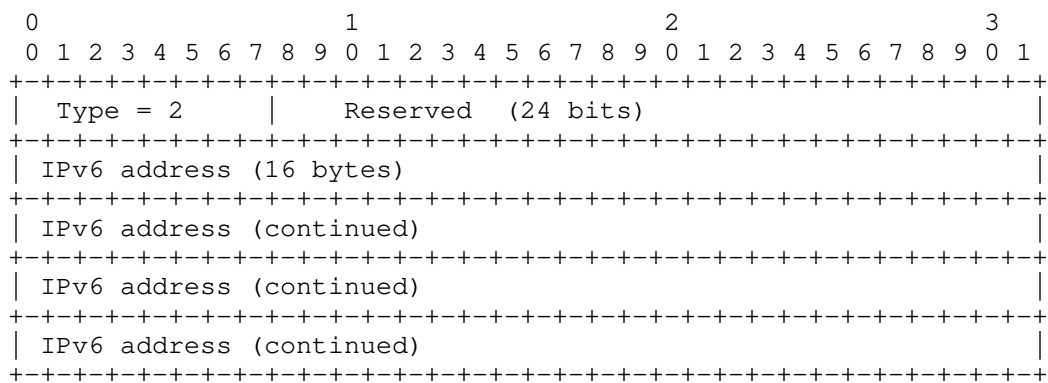
<IPv4 Address> | <IPv6 Address> | <Unnumbered IF ID>

The encoding of each case is as follows:

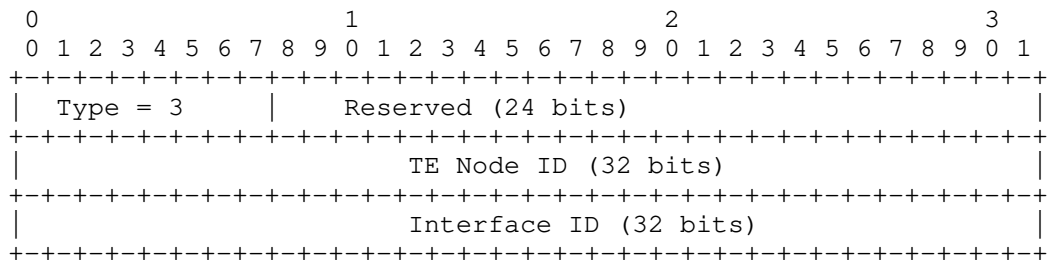
## IPv4 Address Field



## IPv6 Address Field



## Unnumbered Interface ID Address Field



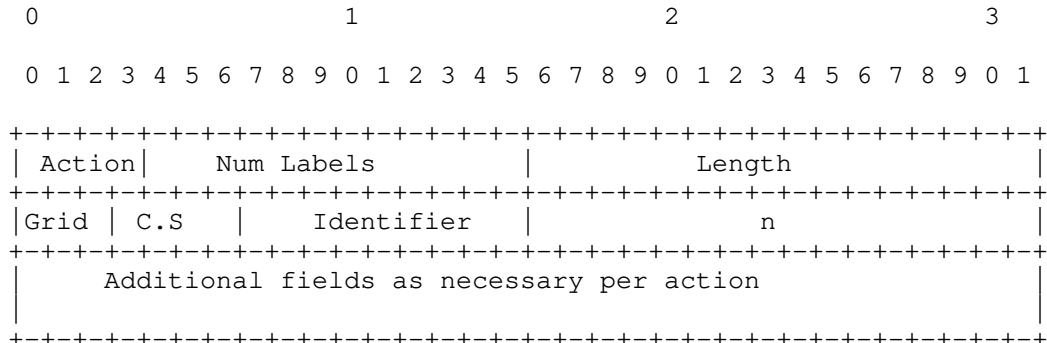
- o Type (8 bits): It indicates the type of the link identifier.

- o Reserved (24 bits): Reserved for future use and SHOULD be zeroed and ignored on receipt.
- o Link Identifier: When Type field is 1, 4-bytes IPv4 address is encoded; when Type field is 2, 16-bytes IPv6 address is encoded; when Type field is 3, a tuple of 4-bytes TE node ID and 4-bytes interface ID is encoded.

The Type field is extensible and matches to the IANA registry created for Link Management Protocol (LMP) [RFC4204] for "TE Link Object Class Type name space": <https://www.iana.org/assignments/lmp-parameters/lmp-parameters.xhtml#lmp-parameters-15>. See Section 8.14 for the request to update the introductory text of the aforementioned registry to note that the values have additional usage for the Link Identifier Type field.

#### 4.3.2. Wavelength Restriction Field

The Wavelength Restriction Field of the Wavelength Restriction Constraint TLV is encoded as a Label Set field as specified in Section 2.6 in [RFC7579] with base label encoded as a 32 bit LSC label, defined in [RFC6205]. The Label Set format is repeated here for convenience, with the base label internal structure included. See [RFC6205] for a description of Grid, C.S, Identifier and n, as well as [RFC7579] for the details of each action.



Action (4 bits):

- 0 - Inclusive List

- 1 - Exclusive List
- 2 - Inclusive Range
- 3 - Exclusive Range
- 4 - Bitmap Set

Num Labels (12 bits): It is generally the number of labels. It has a specific meaning depending on the action value.

Length (16 bits): It is the length in bytes of the entire Wavelength Restriction field.

Identifier (9 bits): The Identifier is always set to 0. If PCC receives the value of the identifier other than 0, it will ignore.

See Sections 2.6.1 - 2.6.3 of [RFC7579] for details on additional field discussion for each action.

#### 4.4. Signal Processing Capability Restrictions

Path computation for WSON includes checking of signal processing capabilities at each interface against requested capability; the PCE MUST have mechanisms to know the signal processing capabilities at each interface, e.g. by means of the Traffic Engineering Database (TED) either via IGP or Network Management System (NMS). Moreover, a PCC should be able to indicate additional restrictions to signal processing compatibility, either on the endpoint or any given link.

The supported signal processing capabilities considered in the RWA Information Model [RFC7446] are:

- o Optical Interface Class List
- o Bit Rate
- o Client Signal

The Bit Rate restriction is already expressed in [PCEP-GMPLS] in the BANDWIDTH object.

In order to support the Optical Interface Class information and the Client Signal information new TLVs are introduced as endpoint-restriction in the END-POINTS type Generalized endpoint:

- o Client Signal TLV
- o Optical Interface Class List TLV

The END-POINTS type generalized endpoint is extended as follows:

```
<endpoint-restriction> ::=  
    <LABEL-REQUEST> <label-restriction-list>
```

```
<label-restriction-list> ::= <label-restriction>  
    [<label-restriction-list>]
```

```
<label-restriction> ::= (<LABEL-SET>|  
    [<Wavelength Restriction Constraint>]  
    [<signal-compatibility-restriction>])
```

Where

```
<signal-compatibility-restriction> ::=  
    [<Optical Interface Class List>] [<Client Signal>]
```

The Wavelength Restriction Constraint TLV is defined in Section 4.3.

A new TLV for the Optical Interface Class List TLV (TBD5) is defined, and the encoding of the value part of the Optical Interface Class List TLV is described in Section 4.1 of [RFC7581].

A new TLV for the Client Signal Information TLV (TBD6) is defined, and the encoding of the value part of the Client Signal Information TLV is described in Section 4.2 of [RFC7581].

#### 4.4.1. Signal Processing Exclusion

The PCC/PCE should be able to exclude particular types of signal processing along the path in order to handle client restriction or multi-domain path computation. [RFC5440] defines how Exclude Route Object (XRO) subobject is used. In this draft, we add two new XRO Signal Processing Exclusion Subobjects.

The first XRO subobject type (TBD9) is the Optical Interface Class List Field defined as follows:

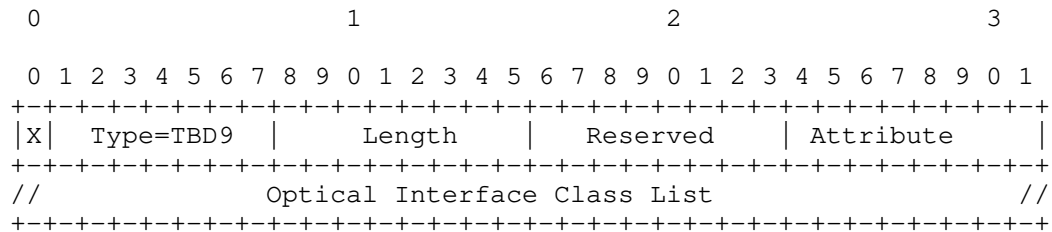


Figure 5 Optical Interface Class List XRO Subobject

Refer to [RFC5521] for the definition of X, Length and Attribute.

Type (7 bits): The Type of the Signaling Processing Exclusion Field. The TLV Type value (TBD9) is to be assigned by the IANA for the Optical Interface Class List XRO Subobject Type.

Reserved bits (8 bits) are for future use and SHOULD be zeroed and ignored on receipt.

The Attribute field (8 bits): [RFC5521] defines several Attribute values; the only permitted Attribute values for this field are 0 (Interface) or 1 (Node).

The Optical Interface Class List is encoded as described in Section 4.1 of [RFC7581].

The second XRO subobject type (TBD10) is the Client Signal Information defined as follows:

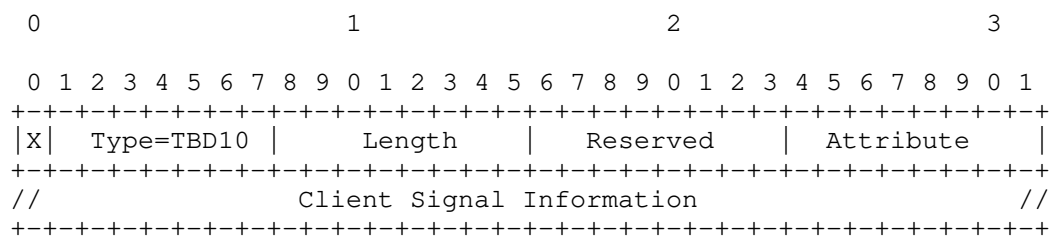




Figure 6 Client Signal Information XRO Subobject

Refer to [RFC5521] for the definition of X, Length and Attribute.

Type (7 bits): The Type of the Signaling Processing Exclusion Field. The TLV Type value (TBD10) is to be assigned by the IANA for the Client Signal Information XRO Subobject Type.

Reserved bits (8 bits) are for future use and SHOULD be zeroed and ignored on receipt.

The Attribute field (8 bits): [RFC5521] defines several Attribute values; the only permitted Attribute values for this field are 0 (Interface) or 1 (Node).

The Client Signal Information is encoded as described in Section 4.2 of [RFC7581].

The XRO needs to support the new Signaling Processing Exclusion XRO Subobject types:

Type	XRO Subobject Type
TBD9	Optical Interface Class List
TBD10	Client Signal Information

#### 4.4.2. Signal Processing Inclusion

Similar to the XRO subobject, the PCC/PCE should be able to include particular types of signal processing along the path in order to handle client restriction or multi-domain path computation. [RFC5440] defines how Include Route Object (IRO) subobject is used. In this draft, we add two new Signal Processing Inclusion Subobjects.

The IRO needs to support the new IRO Subobject types (TBD11 and TBD12) for the PCEP IRO object [RFC5440]:

Type	IRO Subobject Type
------	--------------------

TBD11      Optical Interface Class List

TBD12      Client Signal Information

The encoding of the Signal Processing Inclusion subobjects is similar to Section 4.4.1 where the 'X' field is replaced with 'L' field, all the other fields remains the same. The 'L' field is described in [RFC3209].

## 5. Encoding of a RWA Path Reply

This section provides the encoding of a RWA Path Reply for wavelength allocation request as discussed in Section 4.

### 5.1. Wavelength Allocation TLV

Recall that wavelength allocation can be performed by the PCE by different means:

- (a) By means of Explicit Label Control (ELC) where the PCE allocates which label to use for each interface/node along the path.
- (b) By means of a Label Set where the PCE provides a range of potential labels to allocate by each node along the path.

Option (b) allows distributed label allocation (performed during signaling) to complete wavelength allocation.

The Wavelength Allocation TLV type is TBD4 (See Section 8.4). Note that this TLV is used for both (a) and (b). The TLV data is defined as follows:

```

      0                               1                               2                               3
      0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
|                               Reserved                               |                               Flag                               |M|
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
|                               Link Identifier Field                               |                               |
//                               . . .                               //
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
|                               Allocated Wavelength(s)                               |                               |
//                               . . . .                               //
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+

```

Figure 7 Wavelength Allocation TLV Encoding

- o Reserved (16 bits): Reserved for future use.

- o Flags (16 bits)

One flag bit is allocated as follows:

- . M (Mode): 1 bit

- 0 indicates the allocation is under Explicit Label Control.
    - 1 indicates the allocation is expressed in Label Sets.

IANA is to create a new registry to manage the Flag field (TBD14) of the Wavelength Allocation TLV.

Note that all link identifiers in the same list must be of the same type.

- o Link Identifier: Identifies the interface to which assignment wavelength(s) is applied. See Section 4.3.1. for Link Identifier encoding.

- o Allocated Wavelength(s): Indicates the allocated wavelength(s) to be associated with the Link Identifier. See Section 4.3.2 for encoding details.

This TLV is carried in a PCRep message as an attribute TLV [RFC5420] in the Hop Attribute Subobjects [RFC7570] in the ERO [RFC5440].

## 5.2. Error Indicator

To indicate errors associated with the RWA request, a new Error Type (TBD8) and subsequent error-values are defined as follows for inclusion in the PCEP-ERROR Object:

A new Error-Type (TBD8) and subsequent error-values are defined as follows:

- o Error-Type=TBD8; Error-value=1: if a PCE receives a RWA request and the PCE is not capable of processing the request due to insufficient memory, the PCE MUST send a PCErr message with a PCEP-ERROR Object (Error-Type=TBD8) and an Error-value (Error-value=1). The PCE stops processing the request. The corresponding RWA request MUST be cancelled at the PCC.
- o Error-Type=TBD8; Error-value=2: if a PCE receives a RWA request and the PCE is not capable of RWA computation, the PCE MUST send a PCErr message with a PCEP-ERROR Object (Error-Type=TBD8) and an Error-value (Error-value=2). The PCE stops processing the request. The corresponding RWA computation MUST be cancelled at the PCC.
- o Error-Type=TBD8; Error-value=3: if a PCE receives a RWA request and there are syntactical encoding errors (e.g., not exactly two link identifiers with the range case, unknown identifier types, no matching link for a given identifier, unknown Action value, etc.), the PCE MUST send a PCErr message with a PCEP-ERROR Object (Error-Type=TBD8) and an Error-value (Error-value=3).

### 5.3. NO-PATH Indicator

To communicate the reason(s) for not being able to find RWA for the path request, the NO-PATH object can be used in the corresponding response. The format of the NO-PATH object body is defined in [RFC5440]. The object may contain a NO-PATH-VECTOR TLV to provide additional information about why a path computation has failed.

One new bit flag is defined to be carried in the Flags field in the NO-PATH-VECTOR TLV carried in the NO-PATH Object.

- o Bit TBD7: When set, the PCE indicates no feasible route was found that meets all the constraints (e.g., wavelength restriction, signal compatibility, etc.) associated with RWA.

## 6. Manageability Considerations

Manageability of WSON Routing and Wavelength Assignment (RWA) with PCE must address the following considerations:

### 6.1. Control of Function and Policy

In addition to the parameters already listed in Section 8.1 of [RFC5440], a PCEP implementation SHOULD allow configuration of the following PCEP session parameters on a PCC:

- o The ability to send a WSON RWA request.

In addition to the parameters already listed in Section 8.1 of [RFC5440], a PCEP implementation SHOULD allow configuration of the following PCEP session parameters on a PCE:

- o The support for WSON RWA.
- o A set of WSON RWA specific policies (authorized sender, request rate limiter, etc).

These parameters may be configured as default parameters for any PCEP session the PCEP speaker participates in, or may apply to a specific session with a given PCEP peer or a specific group of sessions with a specific group of PCEP peers.

### 6.2. Liveness Detection and Monitoring

Mechanisms defined in this document do not imply any new liveness detection and monitoring requirements in addition to those already listed in section 8.3 of [RFC5440].

### 6.3. Verifying Correct Operation

Mechanisms defined in this document do not imply any new verification requirements in addition to those already listed in section 8.4 of [RFC5440]

### 6.4. Requirements on Other Protocols and Functional Components

The PCEP Link-State mechanism [PCEP-LS] may be used to advertise WSON RWA path computation capabilities to PCCs.

## 6.5. Impact on Network Operation

Mechanisms defined in this document do not imply any new network operation requirements in addition to those already listed in section 8.6 of [RFC5440].

## 7. Security Considerations

The security considerations discussed in [RFC5440] are relevant for this document, this document does not introduce any new security issues. If an operator wishes to keep private the information distributed by WSON, PCEPS [RFC8253] SHOULD be used.

## 8. IANA Considerations

IANA maintains a registry of PCEP parameters. IANA has made allocations from the sub-registries as described in the following sections.

### 8.1. New PCEP Object: Wavelength Assignment Object

As described in Section 4.1, a new PCEP Object is defined to carry wavelength assignment related constraints. IANA is to allocate the following from "PCEP Objects" sub-registry (<http://www.iana.org/assignments/pcep/pcep.xhtml#pcep-objects>):

Object Class Value	Name	Object Type	Reference
-----			
TBD1	WA	1: Wavelength Assignment	[This.I-D]

### 8.2. WA Object Flag Field

As described in Section 4.1, IANA is to create a registry to manage the Flag field of the WA object. New values are to be assigned by Standards Action [RFC8126]. Each bit should be tracked with the following qualities:

- o Bit number (counting from bit 0 as the most significant bit)
- o Capability description
- o Defining RFC

The following values are defined in this document:

One bit is defined for the WA Object flag in this document:

Codespace of the Flag field (WA Object)

Bit	Description	Reference
-----		
0-14	Unassigned	[This.I-D]
15	Explicit Label Control	[This.I-D]

### 8.3. New PCEP TLV: Wavelength Selection TLV

As described in Sections 4.2, a new PCEP TLV is defined to indicate wavelength selection constraints. IANA is to allocate this new TLV from the "PCEP TLV Type Indicators" subregistry (<http://www.iana.org/assignments/pcep/pcep.xhtml#pcep-tlv-type-indicators>).

Value	Description	Reference
-----		
TBD2	Wavelength Selection	[This.I-D]

### 8.4. New PCEP TLV: Wavelength Restriction Constraint TLV

As described in Sections 4.3, a new PCEP TLV is defined to indicate wavelength restriction constraints. IANA is to allocate this new TLV from the "PCEP TLV Type Indicators" subregistry (<http://www.iana.org/assignments/pcep/pcep.xhtml#pcep-tlv-type-indicators>).

Value	Description	Reference
-----		
TBD3	Wavelength Restriction	[This.I-D]

## Constraint

## 8.5. Wavelength Restriction Constraint TLV Action Values

As described in Section 4.3, IANA is to allocate a new registry to manage the Action values of the Action field in the Wavelength Restriction Constraint TLV. New values are assigned by Standards Action [RFC8126]. Each value should be tracked with the following qualities: value, meaning, and defining RFC. The following values are defined in this document:

Value	Meaning	Reference
-----		
0	Inclusive List	[This.I-D]
1	Inclusive Range	[This.I-D]
2-255	Reserved	[This.I-D]

## 8.6. New PCEP TLV: Wavelength Allocation TLV

As described in Section 5.1, a new PCEP TLV is defined to indicate the allocation of wavelength(s) by the PCE in response to a request by the PCC. IANA is to allocate this new TLV from the "PCEP TLV Type Indicators" subregistry (<http://www.iana.org/assignments/pcep/pcep.xhtml#pcep-tlv-type-indicators>).

Value	Description	Reference
-----		
TBD4	Wavelength Allocation	[This.I-D]

## 8.7. Wavelength Allocation TLV Flag Field

As described in Section 5.1, IANA is to allocate a registry to manage the Flag field of the Wavelength Allocation TLV. New values are to be assigned by Standards Action [RFC8126]. Each bit should be tracked with the following qualities:

- o Bit number (counting from bit 0 as the most significant bit)



- o Capability description
- o Defining RFC

One bit is defined for the Wavelength Allocation flag in this - document:

Codespace of the Flag field (Wavelength Allocation TLV)

Bit	Description	Reference
-----		
0-14	Unassigned	[This.I-D]
15	Wavelength Allocation Mode	[This.I-D]

#### 8.8. New PCEP TLV: Optical Interface Class List TLV

As described in Section 4.4, a new PCEP TLV is defined to indicate the optical interface class list. IANA is to allocate this new TLV from the "PCEP TLV Type Indicators" subregistry (<http://www.iana.org/assignments/pcep/pcep.xhtml#pcep-tlv-type-indicators>).

Value	Description	Reference
-----		
TBD5	Optical Interface Class List	[This.I-D]

#### 8.9. New PCEP TLV: Client Signal TLV

As described in Section 4.4, a new PCEP TLV is defined to indicate the client signal information. IANA is to allocate this new TLV from the "PCEP TLV Type Indicators" subregistry (<http://www.iana.org/assignments/pcep/pcep.xhtml#pcep-tlv-type-indicators>).

Value	Description	Reference
-----		

TBD6                      Client Signal Information    [This.I-D]

#### 8.10. New No-Path Reasons

As described in Section 5.3, a new bit flag are defined to be carried in the Flags field in the NO-PATH-VECTOR TLV carried in the NO-PATH Object. This flag, when set, indicates that no feasible route was found that meets all the RWA constraints (e.g., wavelength restriction, signal compatibility, etc.) associated with a RWA path computation request.

IANA is to allocate this new bit flag from the "PCEP NO-PATH-VECTOR TLV Flag Field" subregistry  
(<http://www.iana.org/assignments/pcep/pcep.xhtml#no-path-vector-tlv>).

Bit	Description	Reference
-----		
TBD7	No RWA constraints met	[This.I-D]

#### 8.11. New Error-Types and Error-Values

As described in Section 5.2, new PCEP error codes are defined for WSON RWA errors. IANA is to allocate from the "PCEP-ERROR Object Error Types and Values" sub-registry  
(<http://www.iana.org/assignments/pcep/pcep.xhtml#pcep-error-object>).

Error-Type	Meaning	Error-Value	Reference
-----			
TBD8	WSON RWA Error	0: Unassigned	[This.I-D]
		1: Insufficient Memory	[This.I-D]
		2: RWA computation Not supported	[This.I-D]

3: Syntactical [This.I-D]  
Encoding error

4-255: Unassigned [This.I-D]

#### 8.12. New Subobjects for the Exclude Route Object

As described in Section 4.4.1, the "PCEP Parameters" registry contains a subregistry "PCEP Objects" with an entry for the Exclude Route Object (XRO). IANA is requested to add further subobjects that can be carried in the XRO as follows:

Subobject	Type	Reference
-----		
TBD9	Optical Interface Class List	[This.I-D]
TBD10	Client Signal Information	[This.I-D]

#### 8.13. New Subobjects for the Include Route Object

As described in Section 4.4.2, the "PCEP Parameters" registry contains a subregistry "PCEP Objects" with an entry for the Include Route Object (IRO). IANA is requested to add further subobjects that can be carried in the IRO as follows:

Subobject	Type	Reference
-----		
TBD11	Optical Interface Class List	[This.I-D]
TBD12	Client Signal Information	[This.I-D]

#### 8.14. Request for Updated Note for LMP TE Link Object Class Type

As discussed in Section 4.3.1, the registry created for Link Management Protocol (LMP) [RFC4204] for "TE Link Object Class Type name space": <https://www.iana.org/assignments/lmp-parameters/lmp-parameters.xhtml#lmp-parameters-15> is requested for the updated

introductory note that the values have additional usage for the Link Identifier Type field.

## 9. Acknowledgments

The authors would like to thank Adrian Farrel, Julien Meuric, Dhruv Dhody and Benjamin Kaduk for many helpful comments that greatly improved the contents of this draft.

## 10. References

### 10.1. Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC3209] D. Awduche, L. Berger, D. Gan, T. Li, V. Srinivasan, G. Swallow, "RSVP-TE: Extensions to RSVP for LSP Tunnels", RFC 3209, December 2001.
- [RFC3630] D. Katz, K. Kompella, D. Yeung, "Traffic Engineering (TE) Extensions to OSPF Version 2", RFC 3630, September 2003.
- [RFC5329] A. Lindem, Ed., "Traffic Engineering Extensions to OSPF Version 3", RFC 5329, September 2008.
- [RFC5440] JP. Vasseur, Ed., JL. Le Roux, Ed., "Path Computation Element (PCE) Communication Protocol (PCEP)", RFC 5440, March 2009.
- [RFC6205] Tomohiro, O. and D. Li, "Generalized Labels for Lambda-Switching Capable Label Switching Routers", RFC 6205, January, 2011.
- [RFC7570] C. Margaria, et al., "Label Switched Path (LSP) Attribute in the Explicit Route Object (ERO)", RFC 7570, July 2015.
- [RFC7579] G. Bernstein and Y. Lee, "General Network Element Constraint Encoding for GMPLS Controlled Networks", RFC 7579, June 2015.

- [RFC7581] G. Bernstein and Y. Lee, "Routing and Wavelength Assignment Information Encoding for Wavelength Switched Optical Networks", RFC7581, June 2015.
- [RFC7689] Bernstein et al., "Signaling Extensions for Wavelength Switched Optical Networks", RFC 7689, November 2015.
- [RFC7688] Y. Lee, and G. Bernstein, "OSPF Enhancement for Signal and Network Element Compatibility for Wavelength Switched Optical Networks", RFC 7688, November 2015.
- [RFC8174] B. Leiba, "Ambiguity of Uppercase vs Lowercase in RFC 2119 Key Words", RFC 8174, May 2017.
- [RFC8253] D. Lopez, O. Gonzalez de Dios, Q. Wu, D. Dhody, "PCEPS: Usage of TLS to Provide a Secure Transport for the Path Computation Element Communication Protocol (PCEP)", RFC 8253, October 2017.
- [PCEP-GMPLS] C. Margaria, et al., "PCEP extensions for GMPLS", draft-ietf-pce-gmpls-pcep-extensions, work in progress.

## 10.2. Informative References

- [RFC3471] Berger, L. (Editor), "Generalized Multi-Protocol Label Switching (GMPLS) Signaling Functional Description", RFC 3471. January 2003.
- [RFC4203] K. Kompella, Ed., Y. Rekhter, Ed., "OSPF Extensions in Support of Generalized Multi-Protocol Label Switching (GMPLS)", RFC 4203, October 2005.
- [RFC4204] J. Lang, Ed., "Link Management Protocol (LMP)", RFC 4204, October 2005.
- [RFC4655] A. Farrel, JP. Vasseur, G. Ash, "A Path Computation Element (PCE)-Based Architecture", RFC 4655, August 2006.
- [RFC5420] Farrel, A. "Encoding of Attributes for MPLS LSP Establishment Using Resource Reservation Protocol Traffic Engineering (RSVP-TE)", RFC5420, February 2009.

- [RFC5440] Vasseur, JP., Ed. and JL. Le Roux, Ed., "Path Computation Element (PCE) communication Protocol", RFC 5440, March 2009. [RFC5521] Oki, E, T. Takeda, and A. Farrel, "Extensions to the Path Computation Element Communication Protocol (PCEP) for Route Exclusions", RFC 5521, April 2009.
- [RFC6163] Lee, Y. and Bernstein, G. (Editors), and W. Imajuku, "Framework for GMPLS and PCE Control of Wavelength Switched Optical Networks", RFC 6163, March 2011.
- [RFC6566] Lee, Y. and Bernstein, G. (Editors), "A Framework for the Control of Wavelength Switched Optical Networks (WSONs) with Impairments", RFC 6566, March 2012.
- [RFC7446] Y. Lee, G. Bernstein, (Editors), "Routing and Wavelength Assignment Information Model for Wavelength Switched Optical Networks", RFC 7446, February 2015.
- [RFC7449] Y. Lee, G. Bernstein, (Editors), "Path Computation Element Communication Protocol (PCEP) Requirements for Wavelength Switched Optical Network (WSON) Routing and Wavelength Assignment", RFC 7449, February 2015.
- [PCEP-LS] Y. Lee, et al., "PCEP Extension for Distribution of Link-State and TE information for Optical Networks", draft-lee-pce-pcep-ls-optical, work in progress.
- [RFC8126] M. Cotton, B. Leiba, T. Narten, "Guidelines for Writing an IANA Considerations Section in RFCs", RFC 8126, June 2017.

## 11. Contributors

Fatai Zhang  
Huawei Technologies  
Email: zhangfatai@huawei.com

Cyril Margaria  
Nokia Siemens Networks  
St Martin Strasse 76  
Munich, 81541  
Germany  
Phone: +49 89 5159 16934  
Email: cyril.margaria@nsn.com

Oscar Gonzalez de Dios  
Telefonica Investigacion y Desarrollo  
C/ Emilio Vargas 6  
Madrid, 28043  
Spain  
Phone: +34 91 3374013  
Email: ogondio@tid.es

Greg Bernstein  
Grotto Networking  
Fremont, CA, USA  
Phone: (510) 573-2237  
Email: gregb@grotto-networking.com

Authors' Addresses

Young Lee, Editor  
Huawei Technologies  
5700 Tennyson Parkway Suite 600  
Plano, TX 75024, USA  
Email: leeyoung@huawei.com

Ramon Casellas, Editor  
CTTC PMT Ed B4 Av. Carl Friedrich Gauss 7  
08860 Castelldefels (Barcelona)  
Spain  
Phone: (34) 936452916  
Email: ramon.casellas@cttc.es





Path Computation Element  
Internet-Draft  
Intended status: Standards Track  
Expires: January 11, 2014

D. Lopez  
O. Gonzalez de Dios  
Telefonica I+D  
July 10, 2013

Secure Transport for PCEP  
draft-lopez-pcp-pceps-00

## Abstract

The Path Computation Element Communication Protocol (PCEP) defines the mechanisms for the communication between a client and a PCE, or among PCEs. This document describe the usage of Transport Layer Security to enhance PCEP security, hence the PCEPS acronym proposed for it. The additional security mechanisms are provided by the transport protocol supporting PCEP, and therefore they do not affect its flexibility and extensibility.

## Status of this Memo

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

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

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

This Internet-Draft will expire on January 11, 2014.

## Copyright Notice

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

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

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

## Table of Contents

1. Introduction . . . . .	3
2. Applying TLS to PCEP . . . . .	3
2.1. TCP ports . . . . .	3
2.2. Connection Establishment . . . . .	4
2.3. Peer Identity . . . . .	5
3. IANA Considerations . . . . .	6
4. Security Considerations . . . . .	6
5. Acknowledgements . . . . .	7
6. References . . . . .	7
6.1. Normative References . . . . .	7
6.2. Informative References . . . . .	8
Authors' Addresses . . . . .	8

## 1. Introduction

PCEP [RFC5440] defines the mechanisms for the communication between a Path Computation Client (PCC) and a Path Computation Element (PCE), or between two PCEs. These interactions include requests and replies that can be critical for a sustainable network operation and adequate resource allocation, and therefore appropriate security becomes a key element in the PCE infrastructure. As the applications of the PCE framework evolves, and more complex service patterns emerge, the definition of a secure mode of operation becomes more relevant.

[RFC5440] analyzes in its section on security considerations the potential threats to PCEP and their consequences, and discusses several mechanisms for protecting PCEP against security attacks, without making a specific recommendation on a particular one or defining their application in depth. Moreover, [RFC6952] remarks the importance of ensuring PCEP communication privacy, especially when PCEP communication endpoints do not reside in the same AS, as the interception of PCEP messages could leak sensitive information related to computed paths and resources.

Among the possible solutions mentioned in these documents, Transport Layer Security (TLS) [RFC5246] provides support for peer authentication, and message encryption and integrity. TLS supports the usage of well-know mechanisms to support key configuration and exchange, and means to perform security checks on the results of PCE discovery procedures ([RFC5088] and [RFC5089]). Since TLS is a security container for the transport of PCEP requests and replies, it will not interfere with the protocol flexibility and extensibility.

This document describes how to apply TLS in securing PCE interactions, including the handshake mechanisms, the methods for peer authentication, and the applicable TLS ciphersuites for data exchange. In the rest of the document we will refer to this usage of TLS as transport for PCEP as either "PCEP over TLS" or "PCEPS".

## 2. Applying TLS to PCEP

### 2.1. TCP ports

The default destination port number for PCEP over TLS is TCP/XXXX.

NOTE: This port has to be agreed and registered as PCEPS with IANA.

## 2.2. Connection Establishment

PCEPS has no notion of negotiating TLS in an established connection. Both peers in the connection need to be preconfigured to use PCEPS for a given endpoint. The connection establishment SHALL follow the following steps:

1. After completing the TCP handshake, immediately negotiate TLS sessions according to [RFC5246]. The following restrictions apply:
  - \* Support for TLS v1.2 [RFC5246] or later is REQUIRED.
  - \* Support for certificate-based mutual authentication is REQUIRED.
  - \* Negotiation of mutual authentication is REQUIRED.
  - \* Negotiation of a ciphersuite providing for integrity protection is REQUIRED.
  - \* Negotiation of a ciphersuite providing for confidentiality is RECOMMENDED.
  - \* Support for and negotiation of compression is OPTIONAL.
  - \* PCEPS implementations MUST, at a minimum, support negotiation of the TLS\_RSA\_WITH\_3DES\_EDE\_CBC\_SHA, and SHOULD support TLS\_RSA\_WITH\_RC4\_128\_SHA and TLS\_RSA\_WITH\_AES\_128\_CBC\_SHA as well. In addition, PCEPS implementations MUST support negotiation of the mandatory-to-implement ciphersuites required by the versions of TLS that they support.
2. Peer authentication can be performed in any of the following two REQUIRED operation models:
  - \* TLS with X.509 certificates using PKIX trust models:
    - + Implementations MUST allow the configuration of a list of trusted Certification Authorities for incoming connections.
    - + Certificate validation MUST include the verification rules as per [RFC5280].
    - + Implementations SHOULD indicate their trusted Certification Authorities (CAs). For TLS 1.2, this is done using [RFC5246], Section 7.4.4, "certificate\_authorities" (server side) and [RFC6066], Section 6 "Trusted CA Indication"

(client side).

- + Peer validation always SHOULD include a check on whether the locally configured expected DNS name or IP address of the server that is contacted matches its presented certificate. DNS names and IP addresses can be contained in the Common Name (CN) or subjectAltName entries. For verification, only one of these entries is to be considered. The following precedence applies: for DNS name validation, subjectAltName:DNS has precedence over CN; for IP address validation, subjectAltName:iPAddr has precedence over CN.
- + NOTE: Consider here whether peer validation MAY be extended by means of the DANE procedures, including its specs as informative references.
- + Implementations MAY allow the configuration of a set of additional properties of the certificate to check for a peer's authorization to communicate (e.g., a set of allowed values in subjectAltName:URI or a set of allowed X509v3 Certificate Policies)
- \* TLS with X.509 certificates using certificate fingerprints: Implementations MUST allow the configuration of a list of trusted certificates, identified via fingerprint of the DER encoded certificate octets. Implementations MUST support SHA-256 as the hash algorithm for the fingerprint.

### 3. Start exchanging PCEP requests and replies.

NOTE: TLS re-negotiation left as an open issue.

#### 2.3. Peer Identity

Depending on the peer authentication method in use, PCEPS supports different operation modes to establish peer's identity and whether it is entitled to perform requests or can be considered authoritative in its replies. PCEPS implementations SHOULD provide mechanisms for associating peer identities with different levels of access and/or authoritativeness, and they MUST provide a mechanism for establish a default level for properly identified peers. Any connection established with a peer that cannot be properly identified SHALL be terminated before any PCEP exchange takes place.

In TLS-X.509 mode using fingerprints, a peer is uniquely identified by the fingerprint of the presented client certificate.

There are numerous trust models in PKIX environments, and it is beyond the scope of this document to define how a particular deployment determines whether a client is trustworthy. Implementations that want to support a wide variety of trust models should expose as many details of the presented certificate to the administrator as possible so that the trust model can be implemented by the administrator. As a suggestion, at least the following parameters of the X.509 client certificate should be exposed:

- o Peer's IP address
- o Peer's FQDN
- o Certificate Fingerprint
- o Issuer
- o Subject
- o All X509v3 Extended Key Usage
- o All X509v3 Subject Alternative Name
- o All X509v3 Certificate Policies

NOTE: Additional procedures enabled by DANE methods are TBD

NOTE: Specific connections with PCE discovery procedures is TBD

### 3. IANA Considerations

NOTE: PCEPS has to be registered as TCP port XXXX.

No new PCEP messages or other objects are defined.

### 4. Security Considerations

Since computational resources required by TLS handshake and ciphersuite are higher than unencrypted TCP, clients connecting to a PCEPS server can more easily create high load conditions and a malicious client might create a Denial-of-Service attack more easily.

Some TLS ciphersuites only provide integrity validation of their payload, and provide no encryption. This specification does not forbid the use of such ciphersuites, but administrators must weight carefully the risk of relevant internal data leakage that can occur

in such a case, as explicitly stated by [RFC6952].

When using certificate fingerprints to identify PCEPS peers, any two certificates that produce the same hash value will be considered the same peer. Therefore, it is important to make sure that the hash function used is cryptographically uncompromised so that attackers are very unlikely to be able to produce a hash collision with a certificate of their choice. This document mandates support for SHA-256, but a later revision may demand support for stronger functions if suitable attacks on it are known.

## 5. Acknowledgements

This specification relies on the analysis and profiling of TLS included in [RFC6614].

## 6. References

### 6.1. Normative References

- [RFC5088] Le Roux, JL., Vasseur, JP., Ikejiri, Y., and R. Zhang, "OSPF Protocol Extensions for Path Computation Element (PCE) Discovery", RFC 5088, January 2008.
- [RFC5089] Le Roux, JL., Vasseur, JP., Ikejiri, Y., and R. Zhang, "IS-IS Protocol Extensions for Path Computation Element (PCE) Discovery", RFC 5089, January 2008.
- [RFC5246] Dierks, T. and E. Rescorla, "The Transport Layer Security (TLS) Protocol Version 1.2", RFC 5246, August 2008.
- [RFC5280] Cooper, D., Santesson, S., Farrell, S., Boeyen, S., Housley, R., and W. Polk, "Internet X.509 Public Key Infrastructure Certificate and Certificate Revocation List (CRL) Profile", RFC 5280, May 2008.
- [RFC5440] Vasseur, JP. and JL. Le Roux, "Path Computation Element (PCE) Communication Protocol (PCEP)", RFC 5440, March 2009.
- [RFC6066] Eastlake, D., "Transport Layer Security (TLS) Extensions: Extension Definitions", RFC 6066, January 2011.



## 6.2. Informative References

- [RFC6614] Winter, S., McCauley, M., Venaas, S., and K. Wierenga, "Transport Layer Security (TLS) Encryption for RADIUS", RFC 6614, May 2012.
- [RFC6952] Jethanandani, M., Patel, K., and L. Zheng, "Analysis of BGP, LDP, PCEP, and MSDP Issues According to the Keying and Authentication for Routing Protocols (KARP) Design Guide", RFC 6952, May 2013.

## Authors' Addresses

Diego R. Lopez  
Telefonica I+D  
Don Ramon de la Cruz, 82  
Madrid, 28006  
Spain

Phone: +34 913 129 041  
Email: diego@tid.es

Oscar Gonzalez de Dios  
Telefonica I+D  
Don Ramon de la Cruz, 82  
Madrid, 28006  
Spain

Phone: +34 913 129 041  
Email: ogondio@tid.es



PCE Working Group  
Internet-Draft  
Intended status: Standards Track  
Expires: January 6, 2014

U. Palle  
D. Dhody  
X. Zhang  
Huawei Technologies  
July 5, 2013

LSP-DB Synchronization between Stateful PCEs  
draft-palle-pce-stateful-pce-lspdb-sync-00

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.

[STATEFUL-PCE] specifies a set of extensions to PCEP to enable stateful control of MPLS-TE and GMPLS Label Switched Paths (LSPs) via PCEP and maintaining of these LSPs at the stateful PCE. This document describes the mechanisms of LSP Database (LSP-DB) synchronization between stateful PCEs.

Status of This Memo

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

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

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

This Internet-Draft will expire on January 6, 2014.

Copyright Notice

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

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

carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

## Table of Contents

1. Introduction . . . . .	3
1.1. Requirements Language . . . . .	3
2. Terminology . . . . .	3
3. Motivation & Use . . . . .	3
4. Functions to Support LSP-DB Synchronization . . . . .	4
5. Architectural Overview . . . . .	5
5.1. LSP-DB Synchronization between Primary and Backup Stateful PCEs . . . . .	5
5.2. LSP-DB Synchronization between Load-Balanced Stateful PCEs . . . . .	6
5.3. Other Considerations . . . . .	8
6. PCEP Messages . . . . .	8
6.1. The PCRpt Message . . . . .	8
6.2. The PCUpd Message . . . . .	8
7. TLVs . . . . .	8
7.1. Stateful PCE Capability TLV . . . . .	8
7.2. PCE Redundancy Group Identifier TLV . . . . .	8
7.3. PCE-CAP-FLAGS sub-TLV . . . . .	9
8. Security Considerations . . . . .	9
9. Manageability Considerations . . . . .	9
9.1. Control of Function and Policy . . . . .	9
9.2. Information and Data Models . . . . .	9
9.3. Liveness Detection and Monitoring . . . . .	9
9.4. Verify Correct Operations . . . . .	10
9.5. Requirements On Other Protocols . . . . .	10
9.6. Impact On Network Operations . . . . .	10
10. IANA Considerations . . . . .	10
10.1. STATEFUL-PCE-CAPABILITY TLV . . . . .	10
10.2. PCE-CAP-FLAGS sub-TLV . . . . .	10
11. Acknowledgments . . . . .	10
12. References . . . . .	11
12.1. Normative References . . . . .	11
12.2. Informative References . . . . .	11
Appendix A. Contributor Addresses . . . . .	11

## 1. Introduction

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

[STATEFUL-PCE] specifies a set of extensions to PCEP to enable stateful control of LSPs in compliance with [RFC4655]. It includes mechanisms for LSP state synchronization between a PCC and a PCE.

This document specifies the mechanisms of LSP-DB synchronization between stateful PCEs in the same domain.

### 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 [RFC5440] and [STATEFUL-PCE].

**LSP-DB:** A database of LSPs that are active in the network as maintained by a stateful PCE.

**Sticky Resources:** The temporarily assigned resources that are allocated to a pending LSP and are provisionally blocked.

## 3. Motivation & Use

"Distributed computation model" ([RFC4655]) refers to a domain or network that may include multiple PCEs where computation of paths is shared among the PCEs, this is further clarified in [PCE-QUESTIONS].

When multiple stateful PCEs are operating in the network, they could be either -

**Primary or Backup PCE:** A backup PCE exists to perform functions in the network, only in the event of a failure of the primary PCE. In this case, all LSPs to be delegated are under primary stateful PCE control while other PCEs in the domain act as backup. The backup PCE should have the same view of LSP-DB as primary stateful PCE. The LSP-DB of a backup PCE can be synchronized via the primary stateful PCE or collected from multiple network nodes (PCC). In case of latter, the backup PCE may face synchronization

issues as described in [PCE-QUESTIONS]. Thus it is suggested that backup PCE can be synchronized via the primary stateful PCE, this mechanism is described in Section 5.1.

**Load-Balanced PCE:** Load-Balanced PCEs share the computation load all the time. One PCE MAY serve a set of PCCs as the primary computation server, and only addresses requests from other PCCs in the event of the failure of some other PCE. Delegated LSPs are thus distributed among stateful PCEs. It is suggested that in this case each load-balanced stateful PCE should build their LSP-DB independently from the network (PCCs) (via mechanism described in [STATEFUL-PCE]) during initial LSP state synchronization and not from other stateful PCEs. But it is important that these load-balanced stateful PCEs needs to be synchronized to have a similar view of pending LSPs and sticky resources, this mechanism is described in Section 5.2.

#### 4. Functions to Support LSP-DB Synchronization

[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).

- o Capability negotiation (E-C,C-E)
- o LSP state synchronization (C-E)
- o LSP update request (E-C)
- o LSP state report (C-E)
- o LSP control delegation (C-E,E-C)
- o Stateful PCE discovery via [STATEFUL-PCE-DISC]

This document extends some of these functions to support LSP-DB synchronization. Some are initiated either from a PCE towards another PCE (E-E) or specifically from primary to backup PCE (PE-BE).

**Capability negotiation (E-E):** both the PCEs must announce during PCEP session establishment that they support PCEP Stateful PCE extensions defined in [STATEFUL-PCE]. It should also declare whether it has primary or backup stateful PCE capability. This is done via Open message.

LSP state synchronization (PE-BE): after the session between the stateful PCEs is initialized, the backup PCE must learn the state of LSPs from the primary PCE. This is done via PCRpt message.

LSP update request (E-E): When a PCE requests modification of attributes of a delegated LSP, this information should also be sent to other PCEs. This is done via PCUpd message. This is needed to synchronize the pending LSPs and sticky resources.

Stateful PCE discovery: PCE can advertise its primary or backup capability via IGP.

## 5. Architectural Overview

LSP-DB synchronization function is defined in section 5.4 of [STATEFUL-PCE] between PCC and PCEs. This document extends the LSP state synchronization between stateful PCEs.

### 5.1. LSP-DB Synchronization between Primary and Backup Stateful PCEs

As shown in Figure 1, PCE1 is the primary stateful PCE and PCE2 is the backup stateful PCE. PCC1 and PCC2 synchronize the LSP-DB with the primary stateful PCE1 after session initialization phase. And primary stateful PCE1 synchronizes LSP-DB with its backup stateful PCE2 after session initialization phase. This is LSP state synchronization as described in Section 4 and uses PCRpt message.

PCC1 & PCC2 delegates LSP1 & LSP2 to the primary PCE1. Whenever there is an update in LSP, PCE1 sends a PCUpd message to corresponding PCC and also to backup PCE2. This is LSP update request as described in Section 4 and uses PCUpd message. This makes sure that the pending LSP changes and sticky resources are backed up. The PCC sends a PCRpt message to the primary PCE, indicating the LSP's status, the primary PCE further synchronizes the state with backup PCEs via PCRpt message.

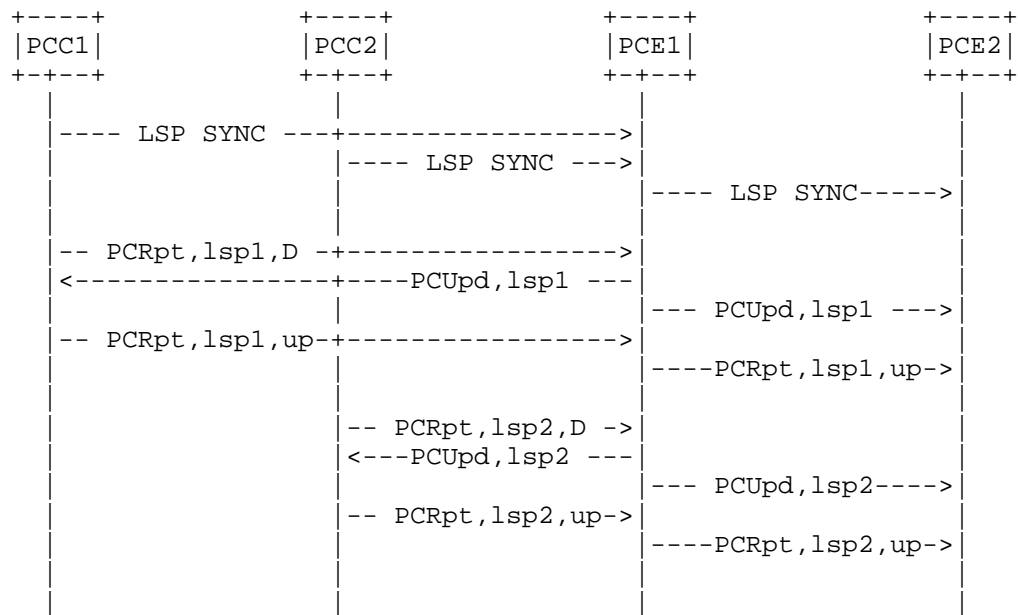


Figure 1: LSP-DB synchronization between primary and backup stateful PCEs

In this case LSP state synchronization (backup) is done via primary stateful PCE. The backup PCE is used only in case the primary PCE fails. The backup PCE SHOULD trust the LSP state as sent by the primary and MAY choose to ignore any state learned from the network (PCCs).

At the time of failure of primary PCE (PCE1), the backup PCE (PCE2) act as a primary. In case of multiple backup PCEs, a selection mechanism (e.g. least IP address among backup PCEs) may be used. When PCE1 recovers from failure, the acting primary PCE (PCE2) should backup using the mechanism as described in this section and restart all its PCEP sessions, thus making sure all PCEP speakers now considers PCE1 as primary.

## 5.2. LSP-DB Synchronization between Load-Balanced Stateful PCEs

As shown in Figure 2, PCE1 and PCE2 are load-balanced stateful PCEs and share the computation load. PCC1 and PCC2 synchronize their LSP-DB with both PCEs after session initialization phase as per [STATEFUL-PCE]. Note that there is no need of LSP-DB state synchronization between PCE1 and PCE2 after session initialization phase as they are load-balanced PCEs and synchronizes the LSP-DB with



the network (PCCs).

PCC1 delegates LSP1 to PCE1. Whenever there is an update in LSP1, PCE1 sends the PCUpd message to PCC1 and other stateful PCEs (PCE2). Similarly, PCC2 delegates LSP2 to PCE2. Whenever there is an update in LSP2, PCE2 sends the PCUpd message to PCC2 and other stateful PCEs (PCE1). This is LSP update request as described in Section 4 and it makes sure that the pending LSP changes and sticky resources are synchronized. The PCC sends an PCRpt message to the all load-balanced PCEs as per [STATEFUL-PCE], indicating the LSP's status.

Note that only the PCUpd message are exchanged between load-balanced PCEs. And the status of the LSPs are received from the network (PCC) via PCRpt message as described in [STATEFUL-PCE].

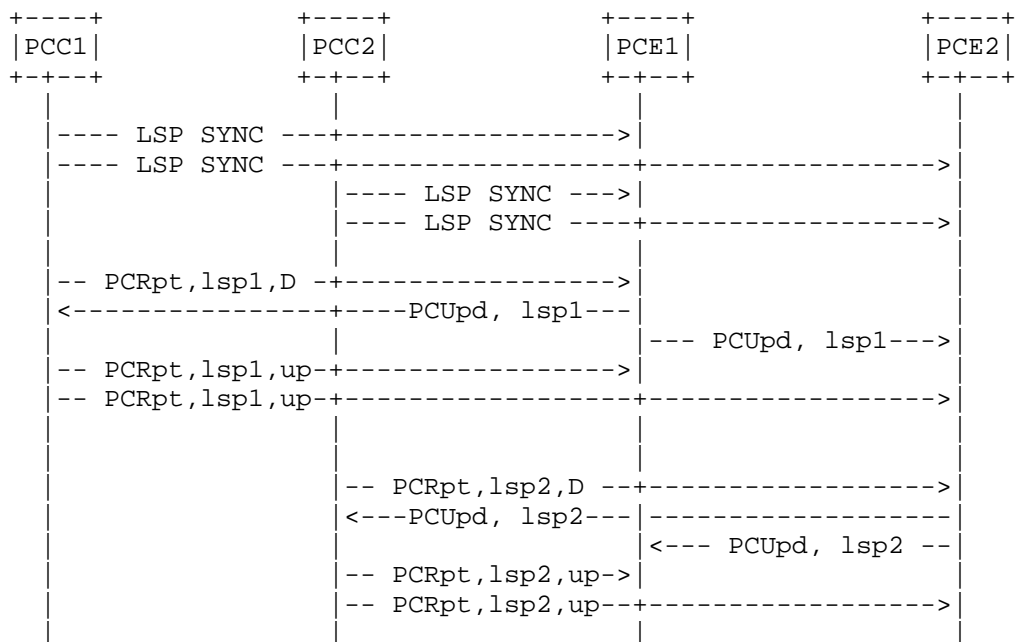


Figure 2: LSP-DB synchronization between load-balanced stateful PCEs

At the time of failure of one of the PCEs (say PCE1), the other PCE (PCE2) may take up the load. When PCE1 recovers from failure, the load can be redistributed again among the PCEs.

### 5.3. Other Considerations

- o This document does not tackle the issue about TED synchronization which is described in detail in [PCE-QUESTIONS].
- o The computation mechanism and how PCE chooses to handle the sticky resources during computation is out of scope.

## 6. PCEP Messages

### 6.1. The PCRpt Message

The format of PCRpt message is defined in [STATEFUL-PCE]. It specifies the PCRpt message is sent from PCC to PCE in reporting the LSP state. This document extends the usage of PCRpt message between primary and backup stateful PCEs for LSP synchronization as described in Section 5.1.

### 6.2. The PCUpd Message

The format of PCUpd Message is defined in [STATEFUL-PCE]. It specifies the PCUpd message is sent from PCE to PCC to request changes in LSP attributes. This document extends the usage of PCUpd message between stateful PCEs for LSP synchronization of pending LSPs and sticky resources as described in Section 5.2. Whenever there is a PCUpd message sent from PCE to PCC, PCE should also send it to other PCEs (backup or load-balanced).

## 7. TLVs

### 7.1. Stateful PCE Capability TLV

As per [STATEFUL-PCE], STATEFUL-PCE-CAPABILITY TLV can be used in the OPEN object for stateful PCE capability negotiation. A stateful PCE must announce during PCEP session establishment that they support PCEP stateful PCE extensions defined in [STATEFUL-PCE]. A new flag is added -

B (BACKUP - 1 bit): if set to 1 by PCE, the PCE should act as a backup. It MAY become an 'acting primary PCE' only in case of failure or unavailability of primary PCE. In case of PCC, this bit has no meaning and is simply ignored.

### 7.2. PCE Redundancy Group Identifier TLV

[STATEFUL-PCE] defines a PREDUNDANCY-GROUP-ID TLV which is a unique identifier of a PCC and carried in OPEN object, [STATEFUL-PCE] also specifies PLSP-ID in LSP object and SYMBOLIC-PATH-NAME TLV which is

used to identify the originating PCC.

To uniquely identify LSP across stateful PCEs, PREDUNDANCY-GROUP-ID TLV MUST be encoded along with LSP object when PCRpt message is sent from primary to backup stateful PCE. This way the backup stateful PCE will also learn the unique identifier for the PCC that does not change.

The existing PREDUNDANCY-GROUP-ID TLV MAYBE encoded in LSP object's optional TLV to identify the originating PCC.

### 7.3. PCE-CAP-FLAGS sub-TLV

[RFC5088] and [RFC5089] describe the mechanism to advertise the PCE Discovery information via OSPF and IS-IS respectively along with processing rules for the sub-TLVs. [STATEFUL-PCE-DISC] further enhances the optional PCE-CAP-FLAGS sub-TLV used to advertise PCE stateful capabilities.

Further a new bit is added -

Bit	Capabilities
TBD	Backup Stateful PCE

If this bit is set to 1, the PCE should act as a backup. It MAY become an 'acting primary PCE' only in case of failure or unavailability of primary PCE.

## 8. Security Considerations

TBD.

## 9. Manageability Considerations

### 9.1. Control of Function and Policy

TBD.

### 9.2. Information and Data Models

TBD.

### 9.3. Liveness Detection and Monitoring

TBD.

#### 9.4. Verify Correct Operations

TBD.

#### 9.5. Requirements On Other Protocols

TBD.

#### 9.6. Impact On Network Operations

TBD.

### 10. IANA Considerations

#### 10.1. STATEFUL-PCE-CAPABILITY TLV

As discussed in Section 7.1, a new STATEFUL-PCE-CAPABILITY TLV Flag Field has been defined. IANA has made the following allocation from the PCEP "STATEFUL-PCE-CAPABILITY TLV Flag Field" sub-registry:

Bit	Description	Reference
TBD	BACKUP	[This I.D.]

#### 10.2. PCE-CAP-FLAGS sub-TLV

As discussed in Section 7.1, a new bit is added, IANA is requested to allocate a new bit in "PCE Capability Flags" registry for backup stateful PCE capability as follows:

Bit	Description	Reference
TBD	BACKUP	[This I.D.]

### 11. Acknowledgments

Thanks to Adrian Farrel and Daniel King for writing [PCE-QUESTIONS].

We would like to thank Avantika Kumar for her useful comments and suggestions.

### 12. References

## 12.1. Normative References

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

## 12.2. Informative References

- [RFC4655] Farrel, A., Vasseur, J., and J. Ash, "A Path Computation Element (PCE)-Based Architecture", RFC 4655, August 2006.
- [RFC5088] Le Roux, JL., Vasseur, JP., Ikejiri, Y., and R. Zhang, "OSPF Protocol Extensions for Path Computation Element (PCE) Discovery", RFC 5088, January 2008.
- [RFC5089] Le Roux, JL., Vasseur, JP., Ikejiri, Y., and R. Zhang, "IS-IS Protocol Extensions for Path Computation Element (PCE) Discovery", RFC 5089, January 2008.
- [RFC5440] Vasseur, JP. and JL. Le Roux, "Path Computation Element (PCE) Communication Protocol (PCEP)", RFC 5440, March 2009.
- [STATEFUL-PCE] Crabbe, E., Medved, J., Minei, I., and R. Varga,, "PCEP Extensions for Stateful PCE (draft-ietf-pce-stateful-pce)", June 2013.
- [PCE-QUESTIONS] Farrel, A. and D. King, "Unanswered Questions in the Path Computation Element Architecture (draft-ietf-pce-questions)", July 2013.
- [STATEFUL-PCE-DISC] Sivabalan, S., Medved, J., and X. Zhang, "IGP Extensions for Stateful PCE Discovery (draft-sivabalan-pce-disco-stateful-01)", April 2013.

## Appendix A. Contributor Addresses

Young Lee  
Huawei  
1700 Alma Drive, Suite 100  
Plano, TX 75075  
US

Phone: +1 972 509 5599 x2240  
Fax: +1 469 229 5397  
EMail: leeyoung@huawei.com

Authors' Addresses

Udayasree Palle  
Huawei Technologies  
Leela Palace  
Bangalore, Karnataka 560008  
INDIA

EMail: udayasree.palle@huawei.com

Dhruv Dhody  
Huawei Technologies  
Leela Palace  
Bangalore, Karnataka 560008  
INDIA

EMail: dhruv.dhody@huawei.com

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

EMail: zhang.xian@huawei.com



Internet Engineering Task Force  
Internet-Draft  
Intended status: Standards Track  
Expires: January 13, 2014

S. Sivabalan  
J. Medved  
Cisco Systems, Inc.  
X. Zhang  
Huawei Technologies  
July 12, 2013

IGP Extensions for Stateful PCE Discovery  
draft-sivabalan-pce-disco-stateful-02

Abstract

When a PCE is a Label Switching Router (LSR) participating in the Interior Gateway Protocol (IGP), or even a server participating in IGP, its presence and path computation capabilities can be advertised using IGP flooding. Such IGP extensions exist for OSPF and ISIS. This document specifies two new PCE capabilities advertised by IGP.

Status of this Memo

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

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

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

This Internet-Draft will expire on January 13, 2014.

Copyright Notice

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

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



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

## Table of Contents

1. Introduction . . . . .	3
1.1. Requirements Language . . . . .	3
2. Terminology . . . . .	3
3. IGP Extensions for Stateful PCE Capabilities . . . . .	4
4. Backward Compatibility . . . . .	5
5. Management Considerations . . . . .	5
6. Security Considerations . . . . .	5
7. IANA Considerations . . . . .	5
8. References . . . . .	5
8.1. Normative References . . . . .	5
8.2. Informative References . . . . .	6
Authors' Addresses . . . . .	6

## 1. Introduction

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

Stateful PCE [I-D.ietf-pce-stateful-pce] specifies a set of extensions to PCEP to enable stateful control of TE LSPs between and across PCEP sessions in compliance with [RFC4657]. It includes mechanisms to effect LSP state synchronization between PCCs and PCEs, delegation of control of LSPs to PCEs, and PCE control of timing and sequence of path computations within and across PCEP sessions. It focuses on a model where LSPs are configured on the PCC and the LSP's path routing and the timing of its setup is delegated to the PCE.

When PCCs are LSRs participating in the IGP (OSPF or IS-IS), and PCEs are either LSRs or servers also participating in the IGP, an effective mechanism for PCE discovery within an IGP routing domain consists of utilizing IGP advertisements. Such extension to OSPF to IS-IS exists in [RFC5088] and [RFC5089], respectively. Currently, the IGP PCE capability does not indicate whether the advertised PCE is stateful. Advertising active and passive stateful PCE capabilities would facilitate a PCC to learn about available stateful PCEs, as well as about a PCE's capability to modify LSP parameters. A PCC could listen to IGP updates, or use other mechanisms that carry IGP information to interested clients, such as BGP-LS ([I-D.ietf-idr-ls-distribution]) where IGP PCE capability advertisements can be carried in the Opaque Prefix Attribute defined in Section 3.3.3.6. This document extends the IGP PCE capability advertisement mechanism to include active and passive stateful PCEs.

### 1.1. Requirements Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC 2119 [RFC2119]

## 2. Terminology

The following terminology is used in this document:

IGP: Interior Gateway Protocol

IS-IS: Intermediate System to Intermediate System

LSR: Label Switching Router

OSPF: Open Shortest Path First

PCC: Path Computation Client

PCE: Path Computation Client

PCEP: Path Computation Client

### 3. IGP Extensions for Stateful PCE Capabilities

The PCE-CAP-FLAGS sub-TLV is an optional sub-TLV used to advertise PCE capabilities. It MAY be present within the PCED sub-TLV carried by OSPF or IS-IS. [RFC5088] and [RFC5089] provide the description and processing rules for this sub-TLV when carried within OSPF and IS-IS, respectively.

The PCE-CAP-FLAGS sub-TLV has the following format:

- o TYPE: 5
- o LENGTH: Multiple of 4
- o VALUE: This contains an array of units of 32 bit flags with the most significant bit as 0. Each bit represents one PCE capability

PCE capability bits are defined in [RFC5088]. This document defines new capability bits for the stateful PCE as follows:

Bit	Capability
11	Active Stateful PCE capability
12	Passive Stateful PCE capability

Note that while active and passive stateful PCE capabilities may be advertised during discovery, PCEP Speakers that wish to use stateful PCEP MUST negotiate stateful PCEP capabilities during PCEP session setup, as specified in Section 7.1.1 in [I-D.ietf-pce-stateful-pce]. A PCC MAY initiate stateful PCEP capability negotiation at PCEP session setup even if it did not receive any IGP PCE capability advertisements.

#### 4. Backward Compatibility

An LSR that does not support the new IGP PCE capability bits specified in this document silently ignores those bits.

IGP extensions defined in this document do not introduce any new interoperability issues.

#### 5. Management Considerations

A configuration option may be provided for advertising and withdrawing Stateful PCE IGP capability on a PCE.

#### 6. Security Considerations

Security considerations described in [RFC5088] are applicable to stateful PCE capabilities. No additional security measures are required.

#### 7. IANA Considerations

IANA is requested to allocate new bits in "PCE Capability Flags" registry for stateful PCE capability as follows:

Bit	Meaning	Reference
11	Active stateful PCE capability	This document
12	Passive stateful PCE capability	This document

#### 8. References

##### 8.1. Normative References

[I-D.ietf-idr-ls-distribution]  
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-03 (work in progress), May 2013.

[I-D.ietf-pce-stateful-pce]  
Crabbe, E., Medved, J., Minei, I., and R. Varga, "PCEP Extensions for Stateful PCE", draft-ietf-pce-stateful-pce-05 (work in progress), July 2013.

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC5088] Le Roux, JL., Vasseur, JP., Ikejiri, Y., and R. Zhang, "OSPF Protocol Extensions for Path Computation Element (PCE) Discovery", RFC 5088, January 2008.
- [RFC5089] Le Roux, JL., Vasseur, JP., Ikejiri, Y., and R. Zhang, "IS-IS Protocol Extensions for Path Computation Element (PCE) Discovery", RFC 5089, January 2008.
- [RFC5440] Vasseur, JP. and JL. Le Roux, "Path Computation Element (PCE) Communication Protocol (PCEP)", RFC 5440, March 2009.

## 8.2. Informative References

- [RFC4657] Ash, J. and J. Le Roux, "Path Computation Element (PCE) Communication Protocol Generic Requirements", RFC 4657, September 2006.

## Authors' Addresses

Siva Sivabalan  
Cisco Systems, Inc.  
2000 Innovation Drive  
Kanata, Ontario K2K 3E8  
Canada

Email: msiva@cisco.com

Jan Medved  
Cisco Systems, Inc.  
170 West Tasman Drive  
San Jose, CA 95134  
USA

Email: jmedved@cisco.com

Xian Zhang  
Huawei Technologies  
F3-5-B R&D Center, Huawei Base Bantian, Longgang District  
Shenzhen, Guangdong 518129  
P.R.China

Email: zhang.xian@huawei.com



Network Working Group  
Internet-Draft  
Intended status: Standards Track  
Expires: January 13, 2014

S. Sivabalan  
J. Medved  
C. Filsfils  
Cisco Systems, Inc.  
E. Crabbe  
Google, Inc.  
R. Raszuk  
NTT I3  
July 12, 2013

PCEP Extensions for Segment Routing  
draft-sivabalan-pce-segment-routing-01.txt

Abstract

Segment Routing (SR) enables any head-end node to select any path without relying on a hop-by-hop signaling technique (e.g., LDP or RSVP-TE). It depends only on "segments" that are advertised by Link-State Interior Gateway Protocols (IGPs). A Segment Routed Path can be derived from a variety of mechanisms, including an IGP Shortest Path Tree (SPT), explicit configuration, or a Path Computation Element (PCE). This document specifies extensions to the Path Computation Element Protocol (PCEP) that allow a stateful PCE to compute and instantiate Traffic Engineering paths, as well as a PCC to request a path subject to certain constraint(s) and optimization criteria in SR networks.

Requirements Language

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

Status of this Memo

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

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

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



This Internet-Draft will expire on January 13, 2014.

#### Copyright Notice

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

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

## Table of Contents

1. Introduction . . . . .	4
2. Terminology . . . . .	5
3. Overview of PCEP Operation in SR Networks . . . . .	6
4. SR-Specific PCEP Message Extensions . . . . .	7
4.1. The PCReq Message . . . . .	7
4.2. The PCRep Message . . . . .	7
4.3. The PCInitiate Message . . . . .	8
4.4. The PCRpt Message . . . . .	8
4.5. The PCUpd Message . . . . .	9
5. Object Formats . . . . .	9
5.1. The OPEN Object . . . . .	9
5.1.1. The SR PCE Capability TLV . . . . .	9
5.2. The RP Object . . . . .	11
5.2.1. The LSP-PATH-TYPE TLV . . . . .	11
5.3. The SR-ERO Object . . . . .	12
5.3.1. The SR-ERO Subobject . . . . .	12
5.3.2. NAI Associated with SID . . . . .	14
5.3.3. SR-ERO Processing . . . . .	15
6. Backward Compatibility . . . . .	15
7. Management Considerations . . . . .	16
7.1. Policy . . . . .	16
7.2. The PCEP Data Model . . . . .	16
8. Security Considerations . . . . .	16
9. IANA Considerations . . . . .	16
9.1. PCEP Objects . . . . .	16
9.1.1. LSP-SIG-TYPE field in the LSP object . . . . .	16
9.2. PCEP-Error Object . . . . .	17
9.3. PCEP TLV Type Indicators . . . . .	17
9.3.1. LSP-PATH-TYPE Indicators . . . . .	17
10. Contributors . . . . .	17
11. Acknowledgements . . . . .	17
12. References . . . . .	18
12.1. Normative References . . . . .	18
12.2. Informative References . . . . .	19
Authors' Addresses . . . . .	19

## 1. Introduction

SR technology leverages the source routing and tunneling paradigms. A source node can choose a path without relying on hop-by-hop signaling protocols such as LDP or RSVP-TE. Each path is specified as a set of "segments" advertised by link-state routing protocols (IS-IS or OSPF). [I-D.filsfils-rtgwg-segment-routing] provides an introduction to SR technology. The corresponding IS-IS and OSPF extensions are specified in [I-D.previdi-isis-segment-routing-extensions] and [I-D.psenak-ospf-segment-routing-extensions], respectively. Two types of segments have been defined; nodal and adjacency segments. A nodal segment represents a path to a node, whereas an adjacency segment represents a specific adjacency to a node. The SR architecture can be applied to MPLS forwarding plane without any change as well as IPv6 forwarding plane with a new type of routing extension header. A Segment Identifier (SID) is a 32-bit value. In the case of the MPLS data plane, an SR path corresponds to an MPLS Label Switching Path (LSP).

A Segment Routed path (SR path) can be derived from an IGP Shortest Path Tree (SPT). Segment Routed Traffic Engineering paths (SR-TE paths) may not follow IGP SPT. Such paths may be chosen by a suitable network planning tool and provisioned on the source node of the SR-TE path.

[RFC5440] describes Path Computation Element Protocol (PCEP) for communication between a Path Computation Client (PCC) and a Path Control Element (PCE) or between one a pair of PCEs. A PCE computes paths for MPLS Traffic Engineering LSPs (MPLS-TE LSPs) based on various constraints and optimization criteria. [I-D.ietf-pce-stateful-pce] specifies extensions to PCEP that allow a stateful PCE to compute and recommend network paths in compliance with [RFC4657] and defines objects and TLVs for MPLS-TE LSPs. Stateful PCEP extensions provide synchronization of LSP state between a PCC and a PCE or between a pair of PCEs, delegation of LSP control, reporting of LSP state from a PCC to a PCE, controlling the setup and path routing of an LSP from a PCE to a PCC. Stateful PCEP extensions are intended for an operational model in which LSPs are configured on the PCC, and control over them is delegated to the PCE.

A mechanism to dynamically instantiate LSPs on a PCC based on the requests from a stateful PCE or a controller using stateful PCE is specified in [I-D.crabbe-pce-pce-initiated-lsp]. Such mechanism is useful in Software Driven Networks (SDN) applications, such as demand engineering, or bandwidth calendaring.

It is possible to use a stateful PCE for computing one or more SR-TE

paths taking into account various constraints and objective functions. Once a path is chosen, the stateful PCE can instantiate an SR-TE path on the PCC using PCEP extensions specified in [I-D.crabbe-pce-pce-initiated-lsp] along with the SR specific PCEP extensions provided in this document. Similarly, a PCC can request an SR path from either stateful or a stateless PCE.

## 2. Terminology

The following terminologies are used in this document:

ERO: Explicit Route Object

IGP: Interior Gateway Protocol

IS-IS: Intermediate System to Intermediate System

LSR: Label Switching Router

MSD: Maximum SID Depth

NAI: Node or Adjacency Identifier

OSPF: Open Shortest Path First

PCC: Path Computation Client

PCE: Path Computation Element

PCEP: Path Computation Element Protocol

RRO: Record Route Object

SID: Segment Identifier

SR: Segment Routing

SR-ERO: Segment Routed Explicit Route Object

SR Path: Segment Routed Path

SR-RRO: Segment Routed Record Route Object

SR-TE Path: Segment Routed Traffic Engineering Path

### 3. Overview of PCEP Operation in SR Networks

In SR networks, an ingress node of an SR path appends all outgoing packets with an SR header consisting of a list of Segment IDs (SIDs). The header has all necessary information to guide the packets from the ingress node to the egress node of the path, and hence there is no need for any signaling protocol. A SID can represent a nodal segment representing a path to a node or adjacency segment representing path over a specific adjacency.

In a PCEP session, path information is carried in the Explicit Route Object (ERO), which consists of a sequence of subobjects. Various types of ERO subobjects have been specified in [RFC3209], [RFC3473], and [RFC3477]. In SR networks, a PCE needs to specify ERO containing SIDs, and a PCC should be capable of processing such ERO. An ERO containing SIDs can be included in the Path Computation LSP Initiate Request message (PCInitiate) defined in [I-D.crabbe-pce-pce-initiated-lsp], as well as in the Path Computation LSP Update Request (PCUpd) and Path Computation LSP State Report (PCRpt) messages defined in Report (PCRpt) messages defined in [I-D.ietf-pce-stateful-pce].

When a PCEP session between a PCC and a PCE is established, both PCEP Speakers exchange information to indicate their ability to support SR-specific functionality. A PCEP session can carry EROs of different types. However, an ERO carrying SIDs MUST NOT include any other form of EROs, i.e., all subobjects within an ERO MUST represent SID. Furthermore, if an SR path is established using SR-ERO, subsequent PCEP Update and Report messages for that path MUST NOT contain other ERO types. This document specifies new error codes to handle these errors. Should the need to change the ERO type arise, the SR path must be deleted and re-created using a new ERO type.

A PCC MAY include an ERO object in a PCRpt message. In SR networks, a PCC MAY learn the SR actual path actually taken by data packets and report that path to a PCE. Methods used by a PCC to learn SR-TE paths are outside the scope of this document.

In summary, this document:

- o Defines a new PCEP capability, new subobjects, a new TLV, and new PCEP error codes
- o Specifies how two PCEP Speakers can establish a PCEP session that can carry segment routing paths
- o Defines the formats of SR-specific PCEP messages in Backus-Naur Format (BNF).

This document specifies SR extensions for the stateless PCE model defined in [RFC5440], as well as for the active stateful and passive stateful PCE models defined in [I-D.ietf-pce-stateful-pce].

#### 4. SR-Specific PCEP Message Extensions

As defined in [RFC5440], a PCEP message consists of a common header followed by a variable length body made up of mandatory and/or optional objects. PCEP messages and their formats for stateless PCE are defined in [RFC5440]. PCEP messages and their formats for stateful PCE are defined in [I-D.ietf-pce-stateful-pce]. Finally, PCEP messages and their formats for PCE-initiated LSP instantiation are defined in [I-D.crabbe-pce-pce-initiated-lsp].

This document defines changes to PCEP messages and their formats required to carry SR-specific information.

##### 4.1. The PCReq Message

This document does not specify any changes to the PCReq message format. This document proposes a new optional TLV carried in the RP Object (Section 5.2.1), which can be used by a PCC to request path computation for one or more SR TE Paths.

##### 4.2. The PCRep Message

This document defines the format of the PCRep message carrying SR TE Paths. The message is sent by a PCE to a PCC in response to a previously received PCReq message, where the PCC requested computation of SR TE Paths. The format of the SR-specific PCRep message is as follows:

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

Where:

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

```
<response> ::= <RP>
                [<NO-PATH>]
                [<path-list>]
```

Where:

```
<path-list> ::= <SR-ERO> [<path-list>]
```

The RP and NO-PATH Objects are defined in [RFC5440]. The <SR-ERO> object contains the SR TE path and is defined in Section 5.3.

#### 4.3. The PCInitiate Message

The format of the PCInitiate message is as follows:

```
<PCInitiate Message> ::= <Common Header>
                           <PCE-initiated-lsp-list>
```

Where:

```
<PCE-initiated-lsp-list> ::=
    <PCE-initiated-lsp-request>[<PCE-initiated-lsp-list>]
```

```
<PCE-initiated-lsp-request> ::= <SRP>
                                <LSP>
                                <SR-ERO>
```

The <SR-ERO> object contains the SR TE path and is defined in Section 5.3. The <LSP> object in the Common Header MUST include the SYMBOLIC-PATH-NAME TLV.

#### 4.4. The PCRpt Message

An SR-specific PCRpt message is sent by a PCC to a PCE to report the current state of an SR TE Path. A PCRpt message can carry more than one LSP State Report, but all LSP State reports in the SR-Specific PCRpt message MUST be for SR TE Paths. A PCC can send an LSP State Report either in response to an LSP Update Request from a PCE, or asynchronously when the state of an SR TE Path changes.

The format of the SR-specific PCRpt message is as follows:

```
<PCRpt Message> ::= <Common Header>
                     <state-report-list>
```

Where:

```
<state-report-list> ::= <state-report>[<state-report-list>]
```

```
<state-report> ::= <SRP>
                  <LSP>
                  <sr-te-path>
```

Where:

```
<sr-te-path> ::= <SR-ERO>
```

The <SR-ERO> object contains the actual SR TE path used by the PCC and is defined in Section 5.3. The actual SR TE Path may be different from the programmed SR TE Path, for example, when the programmed SR TE Path contains loose hops and the PCC must compute the path between loose hops locally.

The <SRP> and <LSP> objects are defined in [I-D.ietf-pce-stateful-pce]. The <LSP> object MUST include the SYMBOLIC-PATH-NAME TLV. The LSP-sig-type filed in the LSP object MUST be set to TBD (Segment Routing).

#### 4.5. The PCUpd Message

An SR-Specific PCUpd message is sent by a PCE to a PCC to update an SR TE Path. A PCUpd message can carry more than one LSP Update Request.

The format of the SR-specific PCUpd message is as follows:

```
<PCUpd Message> ::= <Common Header>
                        <update-request-list>
```

Where:

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

```
<update-request> ::= <SRP>
                        <LSP>
                        <sr-te-path>
```

Where:

```
<sr-te-path> ::= <SR-ERO>
```

The <SR-ERO> object contains the SR TE path computed by the PCE, and is defined in Section 5.3. The <SRP> and <LSP> objects are defined in [I-D.ietf-pce-stateful-pce]. The LSP object MUST include the SYMBOLIC-PATH-NAME TLV. The LSP-sig-type filed in the LSP object MUST be set to TBD (Segment Routing). Note that compared to the RSVP-TE-specific PCUpd message defined in [I-D.ietf-pce-stateful-pce], the path in the SR-specific PCUpd message does not have attributes, only hops specified in the <SR-ERO> object.

## 5. Object Formats

### 5.1. The OPEN Object

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

#### 5.1.1. The SR PCE Capability TLV

The SR-PCE-CAPABILITY TLV is an optional TLV for use in the OPEN Object to negotiate Segment Routing capability on the PCEP session. The format of the SR-PCE-CAPABILITY TLV is shown in the following figure:



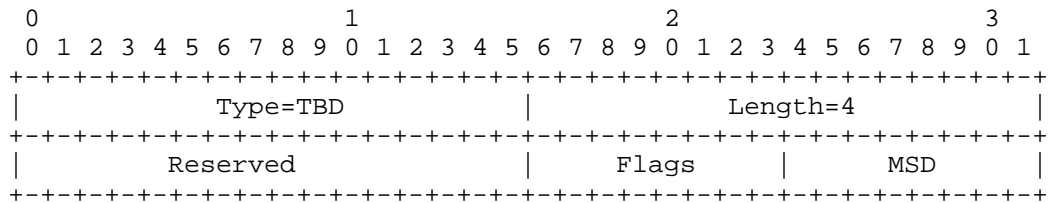


Figure 1: SR-PCE-CAPABILITY TLV format

The code point for the TLV type is to be defined by IANA. The TLV length is 4 octets.

The 32-bit value is formatted as follows. The "Maximum SID Depth" (1 octet) field (MSD) specifies the maximum number of SIDs that a PCC is capable of imposing on a packet. The "Flags" (1 octet) and "Reserved" (2 octets) fields are currently unused, and MUST be set to zero and ignored on receipt.

#### 5.1.1.1. Negotiating SR Capability

The SR capability TLV is contained in the OPEN object. By including the TLV in the OPEN message to a PCE, a PCC indicates its support for SR-TE Paths. By including the TLV in the OPEN message to a PCC, a PCE indicates that it is capable of computing SR-TE paths.

The number of SIDs that can be imposed on a packet depends on PCC's data plane's capability. The default value of MSD is 0 meaning that a PCC does not impose any limitation on the number of SIDs included in any SR-TE path coming from PCE. Once an SR-capable PCEP session is established with a non-default MSD value, the corresponding PCE cannot send SR-TE paths with SIDs exceeding the MSD value. If a PCC needs to modify the MSD value, the PCEP session must be closed and re-established with the new MSD value. If a PCEP session is established with a non-default MSD value, and the PCC receives an SR-TE path containing more SIDs than specified in the MSD value, the PCC MUST send out a PCERR message with Error-Type 10 (Reception of an invalid object) and Error-value 3 (Unsupported number of Segment ERO).

The SR Capability TLV is meaningful only in the OPEN message sent from a PCC to a PCE. As such, a PCE does not need to set MSD value in outbound message to a PCC. Similarly, an MSD value received by a PCC is ignored. If there are multiple SR capability TLVs, only the first TLV is processed.

All bits in the Reserved and Flags fields SHOULD be set to 0 on

outbound OPEN messages, and MUST be ignored on inbound OPEN messages.

## 5.2. The RP Object

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

### 5.2.1. The LSP-PATH-TYPE TLV

A PCC can simultaneously support both RSVP-TE signaled MPLS LSPs as well as SR-TE paths. In this case, the PCC needs to query and receive paths specified with RSVP-TE EROs (defined in [RFC5440]) and paths specified with SR-EROS (defined in this document). Thus, there is a need for a PCC to identify the ERO type that it wishes to receive from a PCE.

This document defines a new optional TLV called "LSP-PATH-TYPE" that MAY be included in the RP object (defined in [RFC5440]) on a PCReq message from a PCC to a PCE. The format of this TLV is shown in the following figure:

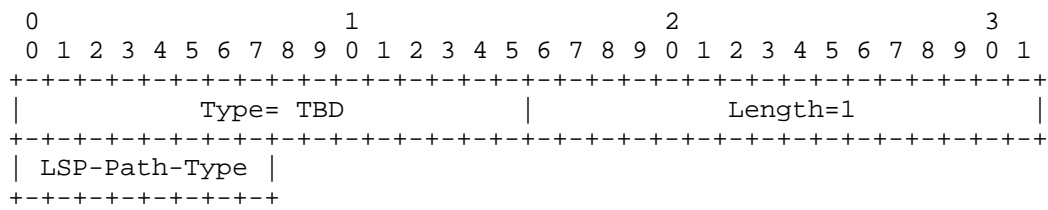


Figure 2: LSP-PATH-TYPE TLV format

The type of the TLV is to be defined by IANA. The TLV length is 1 octets.

The 8-bit value contains the Path-Type (PT). The following values for Path Type are defined:

- o PT = 0: Requested path is to be used with RSVP-TE signaling (default).
- o PT = 1: SR-TE path is requested.

If this TLV is not included in the PCReq message, the default Path Type of 0 is assumed, otherwise the path type specified in the TLV is used. An RP object SHOULD carry no more than one LSP-PATH-TYPE TLV, only the first is used if several are present, the others are ignored.

### 5.3. The SR-ERO Object

An SR-TE path consists of one or more SID(s) where each SID is associated with the identifier that represents the node or adjacency corresponding to the SID. This identifier is referred to as the 'Node or Adjacency Identifier' (NAI). As described later, a NAI can be represented in various formats (e.g., IPv4 address, IPv6 address, etc). Furthermore, a NAI is used only for troubleshooting purposes, and MUST not be used to replace or modify any fields in a data packet header. An SR-ERO object consists of one or more ERO subobjects described in the following section.

#### 5.3.1. The SR-ERO Subobject

An SR-ERO subobject consists of a 32-bit header followed by the SID and the NAI associated with the SID. The SID is a 32-bit number. The size of the NAI depends on its respective type, as described in the following sections.

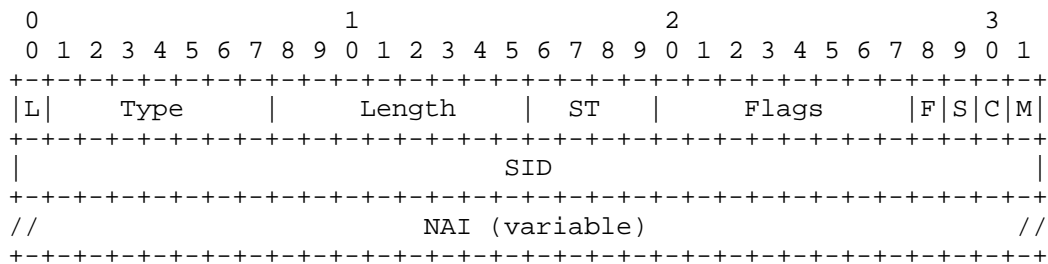


Figure 3: SR-ERO Subobject format

The fields in the ERO Subobject are as follows:

The 'L' Flag indicates whether the subobject represents a loose-hop in the explicit route [RFC3209]. If this flag is unset, a PCC MUST not overwrite the SID value present in the SR-ERO subobject. Otherwise, a PCC MAY expand or replace one or more SID value(s) in the received SR-ERO based on its local policy.

Type is the type of the SR-ERO Subobject. This document defines the SR-ERO Subobject type. A new code point will be requested for the SR-ERO Subobject from IANA.

Length contains the total length of the subobject in octets, including the L, Type and Length fields. Length MUST be at least 4, and MUST be a multiple of 4.

SID Type (ST) indicates the type of information associated with the SID contained in the object body. The SID-Type values are described later in this document.

Flags is used to carry any additional information pertaining to SID. Currently, the following flag bits are defined:

- \* M: When this bit is set, the SID value represents an MPLS label stack entry as specified in [RFC5462] where only the label value is specified by the PCE. Other fields (TC, S, and TTL) fields MUST be considered invalid, and PCC MUST set these fields according to its local policy and MPLS forwarding rules.
- \* C: When this bit as well as the M bit are set, then the SID value represents an MPLS label stack entry as specified in [RFC5462], where all the entry's fields (Label, TC, S, and TTL) are specified by the PCE. However, a PCC MAY choose to override TC, S, and TTL values according its local policy and MPLS forwarding rules.
- \* S: When this bit is set, the SID value in the subobject body is null. In this case, the PCC is responsible for choosing the SID value, e.g., by looking up its Traffic Engineering Database (TED) using node/adjacency identifier in the subobject body.
- \* F: When this bit is set, the NAI value in the subobject body is null.

SID is the Segment Identifier.

NAI contains the NAI associated with the SID. Depending on the value of ST, the NAI can have different format as described in the following section.

## 5.3.2. NAI Associated with SID

This document defines the following NAIs:

'IPv4 Node ID' is specified as an IPv4 address. In this case, ST and Length are 1 and 12 respectively.

'IPv6 Node ID' is specified as an IPv6 address. In this case, ST and Length are 2 and 24 respectively.

'IPv4 Adjacency' is specified as a pair of IPv4 addresses. In this case, ST and Length are 3 and 16, respectively, and the format of the NAI is shown in the following figure:

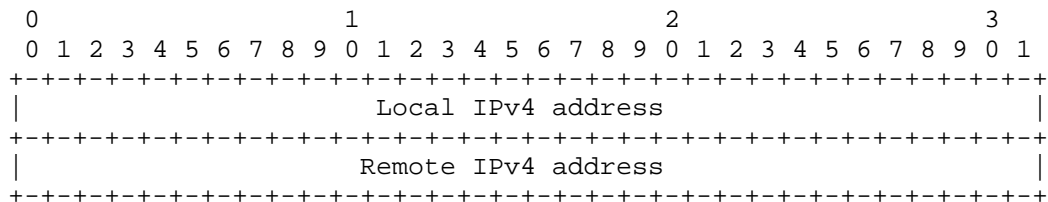


Figure 4: NAI for IPv4 Adjacency

'IPv6 Adjacency' is specified as a pair of IPv6 addresses. In this case, ST and Length are 4 and 40 respectively, and the format of the NAI is shown in the following figure:

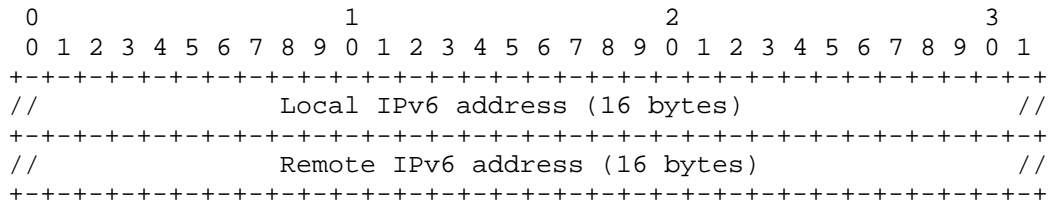


Figure 5: NAI for IPv6 adjacency

'Unnumbered Adjacency with IPv4 NodeIDs' is specified as a pair of Node ID / Interface ID tuples. In this case, ST and Length are 5 and 24 respectively, and the format of the NAI is shown in the following figure:

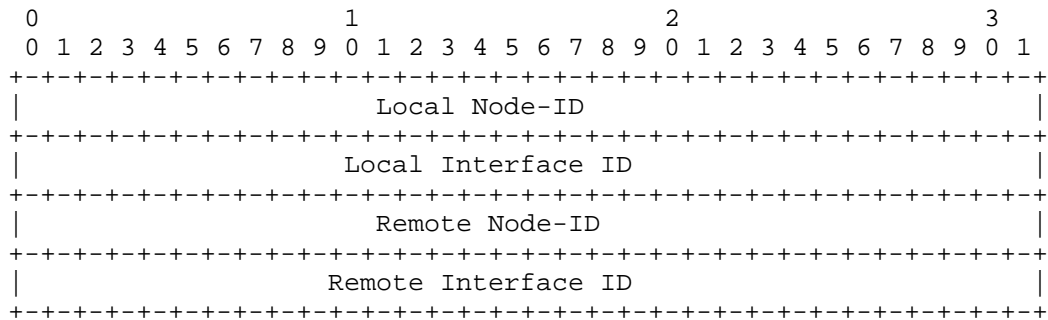


Figure 6: NAI for Unnumbered adjacency with IPv4 Node IDs

We are yet to decide if another SID subobject is required for unnumbered adjacency with 128 bit node ID.

### 5.3.3. SR-ERO Processing

A PCEP Speaker that does not recognize the new ERO subobject in the PCCreate, PCUpd or PCRpt message MUST reject the entire PCEP message and MUST send a PCE error message with Error-Type=3 ("Unknown Object") and Error-Value=2 ("Unrecognized object Type") or Error-Type=4 ("Not supported object") and Error-Value=2 ("Not supported object Type"), defined in [RFC5440].

When the SID represents an MPLS label (i.e. the M bit is set), its value (20 most significant bits) MUST be larger than 15, unless it is special purpose label, such as an Entropy Label Indicator (ELI) or an Entropy Label (EL). If a PCEP Speaker receives a label ERO subobject with an invalid value, it MUST send the PCE error message with Error-Type = "Reception of an invalid object" and Error-Value = "Bad label value". If both M and C bits of an ERO subobject are set, and if a PCEP Speaker finds erroneous setting in one or more of TC, S, and TTL fields, it MUST send a PCE error with Error-Type = "Reception of an invalid object" and Error-Value = "Bad label format".

If a PCC receives a stack of SR-ERO subobjects, and the number of stack exceeds the maximum number of SIDs that the PCC can impose on the packet, it MAY send a PCE error with Error-Type = "Reception of an invalid object" and Error-Value = "Unsupported number of Segment ERO subobjects".

## 6. Backward Compatibility

An LSR that does not support the SR PCEP capability negotiation cannot recognize the SR-ERO subobjects. As such, it shall send a

PCEP error with Error-Type = 4 (Not supported object) and Error-Value = 2 (Not supported object Type) as per [RFC5440].

## 7. Management Considerations

### 7.1. Policy

PCEP implementation:

- o Can enable SR-PCEP capability either by default or via explicit configuration.
- o May generate PCEP error due to unsupported number of SR-ERO subobjects either by default or via explicit configuration.

### 7.2. The PCEP Data Model

A PCEP MIB module is defined in [I-D.ietf-pce-pcep-mib] needs be extended to cover additional functionality provided by [RFC5440] and [I-D.crabbe-pce-pce-initiated-lsp]. Such extension will cover the new functionality specified in this document.

## 8. Security Considerations

The security considerations described in [RFC5440] and [I-D.crabbe-pce-pce-initiated-lsp] are applicable to this specification. No additional security measure is required.

## 9. IANA Considerations

### 9.1. PCEP Objects

IANA is requested to allocate a ERO subobject type (recommended value = 5) for the SR-ERO subobject.

#### 9.1.1. LSP-SIG-TYPE field in the LSP object

This document requests that a new value is allocated for Segment Routing LSP Type in the LSP-SIG\_TYPE registry.

Value	Meaning
1	Segment Routing

## 9.2. PCEP-Error Object

This document defines new Error-Type and Error-Value for the following new conditions:

Error-Type	Meaning
10	Reception of an invalid object
Error-value=2:	Bad label value
Error-value=3:	Unsupported number of Segment ERO subobjects

## 9.3. PCEP TLV Type Indicators

This document defines the following new PCEP TLVs:

Value	Meaning	Reference
26	SR-PCE-CAPABILITY	This document
27	LSP-PATH-TYPE	This document

### 9.3.1. LSP-PATH-TYPE Indicators

This document requests that a registry is created to manage the value of the LSP-Path-Type field in the LSP object, which defines the technology of the LSP path.

Value	Meaning
0	RSVP
1	Segment Routing

## 10. Contributors

The following people contributed to this document:

- Lakshmi Sharma (Cisco Systems)

## 11. Acknowledgements

We'd like to thank Ina Minei and George Swallow for valuable comments.

## 12. References



## 12.1. Normative References

- [I-D.crabbe-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-crabbe-pce-pce-initiated-lsp-01 (work in progress), April 2013.
- [I-D.filsfils-rtgwg-segment-routing]  
Filsfils, C., Previdi, S., Bashandy, A., Decraene, B., Litkowski, S., Horneffer, M., Milojevic, I., Shakir, R., Ytti, S., Henderickx, W., Tantsura, J., and E. Crabbe, "Segment Routing Architecture", draft-filsfils-rtgwg-segment-routing-00 (work in progress), June 2013.
- [I-D.ietf-pce-pcep-mib]  
Koushik, K., Stephan, E., Zhao, Q., King, D., and J. Hardwick, "PCE communication protocol (PCEP) Management Information Base", draft-ietf-pce-pcep-mib-04 (work in progress), February 2013.
- [I-D.ietf-pce-stateful-pce]  
Crabbe, E., Medved, J., Minei, I., and R. Varga, "PCEP Extensions for Stateful PCE", draft-ietf-pce-stateful-pce-05 (work in progress), July 2013.
- [I-D.previdi-isis-segment-routing-extensions]  
Previdi, S., Filsfils, C., Bashandy, A., Gredler, H., Decraene, B., Litkowski, S., Geib, R., Milojevic, I., Shakir, R., Ytti, S., Henderickx, W., and J. Tantsura, "IS-IS Extensions for Segment Routing", draft-previdi-isis-segment-routing-extensions-01 (work in progress), July 2013.
- [I-D.psenak-ospf-segment-routing-extensions]  
Psenak, P., Previdi, S., Filsfils, C., Gredler, H., and R. Shakir, "OSPF Extensions for Segment Routing", draft-psenak-ospf-segment-routing-extensions-01 (work in progress), July 2013.
- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC5440] Vasseur, JP. and JL. Le Roux, "Path Computation Element (PCE) Communication Protocol (PCEP)", RFC 5440, March 2009.

- [RFC5462] Andersson, L. and R. Asati, "Multiprotocol Label Switching (MPLS) Label Stack Entry: "EXP" Field Renamed to "Traffic Class" Field", RFC 5462, February 2009.

## 12.2. Informative References

- [RFC3209] Awduche, D., Berger, L., Gan, D., Li, T., Srinivasan, V., and G. Swallow, "RSVP-TE: Extensions to RSVP for LSP Tunnels", RFC 3209, December 2001.
- [RFC3473] Berger, L., "Generalized Multi-Protocol Label Switching (GMPLS) Signaling Resource ReserVation Protocol-Traffic Engineering (RSVP-TE) Extensions", RFC 3473, January 2003.
- [RFC3477] Kompella, K. and Y. Rekhter, "Signalling Unnumbered Links in Resource ReSerVation Protocol - Traffic Engineering (RSVP-TE)", RFC 3477, January 2003.
- [RFC4657] Ash, J. and J. Le Roux, "Path Computation Element (PCE) Communication Protocol Generic Requirements", RFC 4657, September 2006.

## Authors' Addresses

Siva Sivabalan  
Cisco Systems, Inc.  
2000 Innovation Drive  
Kanata, Ontario K2K 3E8  
Canada

Email: msiva@cisco.com

Jan Medved  
Cisco Systems, Inc.  
170 West Tasman Dr.  
San Jose, CA 95134  
US

Email: jmedved@cisco.com

Clarence Filsfils  
Cisco Systems, Inc.  
Pegasus Parc  
De kleetlaan 6a, DIEGEM BRABANT 1831  
BELGIUM

Email: cfilsfil@cisco.com

Edward Crabbe  
Google, Inc.  
1600 Amphitheatre Parkway  
Mountain View, CA 94043  
US

Email: edward.crabbe@gmail.com

Robert Raszuk  
NTT I3  
101 S. Ellsworth Ave  
San Mateo, CA 94401  
US

Email: robert@raszuk.net



Network Working Group  
Internet-Draft  
Intended status: Standards Track  
Expires: January 16, 2014

Y. Tanaka  
Y. Kamite  
NTT Communications  
I. Minei  
Juniper Networks, Inc.  
Jul 15, 2013

Stateful PCE Extensions for Data Plane Switchover and Balancing  
draft-tanaka-pce-stateful-pce-data-ctrl-00

Abstract

Stateful PCE (Path Computation Element) and its corresponding protocol extensions provide a mechanism that enables PCE to do stateful control of MPLS Traffic Engineering Label Switched Paths (TE LSP). One application that stateful PCE can realize is data traffic reoptimization. Data traffic traversed in a LSP can be switched to another PCE-initiated LSP. Moreover, data traffic can also be balanced to multiple PCE-initiated LSPs using stateful PCE.

This document specifies the extensions to Path Computation Element Protocol (PCEP) that allow a stateful PCE to do switchover and balancing of data traffic with PCE-initiated LSPs. This document also specifies the usage and handling of stateful PCEP (PCE Communication Protocol) messages and the expected behavior of PCC as the RSVP-TE headend.

Status of this Memo

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

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

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

This Internet-Draft will expire on January 16, 2014.

Copyright Notice

Copyright (c) 2013 IETF Trust and the persons identified as the

document authors. All rights reserved.

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

## Table of Contents

1. Introduction . . . . .	3
2. Conventions used in this document . . . . .	4
3. Terminology . . . . .	4
4. PCEP Operation for Data Switchover and Balancing . . . . .	4
5. TLVs in LSP Objects . . . . .	6
5.1. ASSOCIATION-GROUP TLV in LSP Objects . . . . .	6
5.2. DATA-CONTROL TLV in LSP Objects . . . . .	7
6. Operation Examples . . . . .	9
6.1. Data switchover operation (100:0 => 0:100) . . . . .	9
6.2. Load balancing operation (100:0 => 50:50) . . . . .	11
7. IANA Considerations . . . . .	12
7.1. PCEP TLV Indicators . . . . .	12
7.2. PCEP Error Objects . . . . .	12
8. Security Considerations . . . . .	13
9. Acknowledgments . . . . .	13
10. References . . . . .	13
10.1. Normative References . . . . .	13
10.2. Informative References . . . . .	14
Authors' Addresses . . . . .	14

## 1. Introduction

[I-D.ietf-pce-stateful-pce] describes the stateful Path Computation Elements(PCE). Stateful PCE defines the extensions to PCEP to enable stateful control of LSPs between and across PCEP sessions, and it also describes mechanisms to effect LSP state synchronization between PCCs and PCEs, and PCE control of timing and sequence of path computations within and across PCEP sessions. A PCE can update LSP settings (such as bandwidth, priority) using update messages called PCUpd.

[I-D.crabbe-pce-pce-initiated-lsp] defines the extensions to PCEP to allow a PCE to create new LSPs (PCE-Initiated LSP). Before these extensions, the LSP egress point had to be preconfigured at the head end Label Edge Router (LER), the LSP would be set up with default parameters and then its settings (e.g., initial bandwidth, priority) could be modified via PCUpd messages. The extensions for PCE-initiated LSPs eliminate the need for preconfiguration, and allow more flexible operation. Stateful-PCE with LSP instantiation is attracting attention as an enabler for Software Defined Networking (SDN) operation of MPLS networks.

In SDN, it is highly expected to support intelligent and interactive control of the traffic amount of the network by means of a logically-centralized controller. Optimizing the path and bandwidth of MPLS-TE LSP by using stateful PCE is a leading use case of SDN applications. A PCE is able to calculate an optimized route from the topology and bandwidth information in the TED and the LSP state database and it can integrate with a controller that takes into account additional information such as historical trending and service orders in order to trigger PCE action. For example, when data traffic on a LSP counts plenty of bandwidth utilization and if there is no capacity left in the currently signaled path (i.e., no remaining bandwidth of links), a PCE is able to update the existing LSP's parameters (PCE-updated LSP) or create a totally new LSP (PCE-initiated LSP).

The former method is oriented for keeping the existing instance of LSP tunnel, so it does not change a pair of source and destination. Meanwhile, the latter method is oriented for adding a new instance of LSP tunnel.

Specifically regarding the latter method, PCE-initiated LSP, there are some operational scenarios in the network: one is that PCE creates a new LSP that have alternate route with increased-bandwidth LSP and performs switchover from old LSP. Another is that PCE creates one or more additional LSPs and performs load balancing of data traffic. Today, however, there is no detailed procedure specified as to how to control data traffic switching from an old LSP

to new PCE-Initiated LSP(s).

This document specifies the procedures that stateful PCE controls data traffic switchover and load balancing with multiple PCE-Initiated LSPs. This document also specifies the usage and handling of stateful PCEP ( PCE Communication Protocol) messages and the expected behavior of PCC as an RSVP-TE headend.

## 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].

## 3. Terminology

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

This document uses the following terms defined in [I-D.ietf-pce-stateful-pce]: Stateful PCE, LSP State Request, LSP Update Request.

This document uses the following terms defined in [I-D.crabbe-pce-pce-initiated-lsp]: LSP Create Request message.

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

## 4. PCEP Operation for Data Switchover and Balancing

There are two typical operations for explaining the functionality of data switchover and balancing. One is whole data switchover, where a PCC switches all data traffic from one LSP tunnel to another. The other is load balancing of multi-pathing LSP tunnels, where a PCC (headend) balances data traffic among two tunnels equally (fifty percent each, for instance). Both operational cases are completed by the messaging over a single protocol, PCEP, so this is a simple and straightforward solution for MPLS networks.

Support of the procedures listed in this document is negotiated at session init time. The capability negotiation will be speleld out in a future version of this document.

Data switchover and balancing for an MPLS-TE LSP is available once a



PCEP session is established and then a PCC delegates its LSPs to a PCE.

First step is LSP creation. In this step, a PCE sends as many PCInitiate messages as PCE-Initiated LSP as it demands. Once the PCC receives them and successfully establishes PCE-Initiated LSPs, it sends PCRpt messages in reply to the PCInitiate messages and delegates the newly established LSP to the PCE. Message formats and behaviors of the PCC and the PCE are described in detail in [I-D.crabbe-pce-pce-initiated-lsp].

Second step is LSP association. After the PCE-Initiated LSP successfully established and delegated the PCE sends a PCUpd message that contains the ASSOCIATION-GROUP TLV in the LSP Object in order to assemble the members of an association group of LSPs to take over the traffic. Once a PCC receives the PCUpd message with ASSOCIATION-GROUP TLV, the PCC sends back a PCRpt message that contains the ASSOCIATION-GROUP TLV with current operational status.

The option of specifying the association at LSP instantiation time (as part of the PCInitiate message) will be evaluated in a future version of this document.

Third step is executing the data switchover and/or load balancing. In this step, the PCE sends a single PCUpd message which updates the operational status of the LSP from "up and carrying traffic" to just "up". This Update request message for data plane switchover/balancing execution MUST contain DATA-CONTROL TLV in LSP Object. The associated group of traffic origin and that of target to take over the traffic are listed in the DATA-CONTROL TLV. The PCC (LSP headend) load-balances between LSPs in the same association group based on their respective bandwidths. If one of the LSPs does not come up, the traffic would load balance correctly over the others. The switchover case is supported since there will be an association of a single LSP, so that LSP will get hundred percent of data traffic.

The PCC MUST send a PCRpt message to the PCE in order to notify of the result of the data switchover/balancing. The PCRpt message MUST have the DATA-CONTROL TLV that indicates the actual assigned percentages of each member of association group after the execution of the data switchover/balancing operation. The LSP object in the PCRpt will have the reserved PLSP-ID of 0.

The final step is the deletion of old LSP. It is OPTIONAL to carry out this step. The PCE sends PCInitiate message requesting deletion of the LSP that does not carry data traffic anymore after data switchover/balancing execution. Once the PCC tears down the LSP, a

PCRpt message MUST be sent from the PCC to the PCE in order to notify that the LSP is no longer used and return delegation.

Note that, both RSVP-TE [RFC3209] Tunnel-ID and LSP-ID for PCE-Initiated LSP signaling is not allocated by a PCE. A PCC locally assigns those IDs that are related to RSVP-TE parameters. Therefore, the operations of data switchover and balancing specified in this document is the traffic control procedure across multiple RSVP-te Tunnels (i.e., different Tunnel instances). Data switchover method across LSPs within a single RSVP-TE Tunnel, which is the switchover in the middle of make-before-break reoptimization, is covered by [I-D.tanaka-pce-stateful-pce-mbb].

## 5. TLVs in LSP Objects

### 5.1. ASSOCIATION-GROUP TLV in LSP Objects

This section defines ASSOCIATION-GROUP TLV in LSP Objects. ASSOCIATION-GROUP TLV is used in the LSP Object in PCUpd messages when a PCE creates association group of LSPs on a PCC.

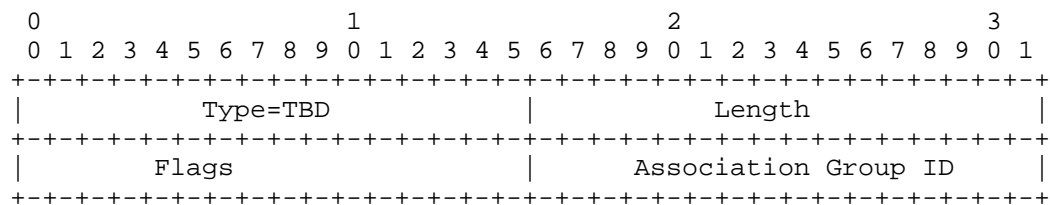


Figure 1: ASSOCIATION-GROUP TLV format

#### Flags and fields

Association Group ID - 8 bit: This field specifies a identifier of association group of LSPs. The IDs are assigned by a PCE, however a PCC (RSVP headend) owns the association group. 0x0000 and 0xFFFF is reserved for special use.

Flags - 8 bit: None defined. MUST be set to zero.

An association group is a group of LSPs that is referenced by a single identifier, by both the PCE and PCC. This number is significant in the context of a single PCEP session. An association group may have one or more LSPs. Association groups with zero

members are removed and the id can be reused. The PCE is the entity managing association, and this is considered PCE state that will be cleaned up when the State Timeout Interval expires. An LSP can be associated with a single association group. Extensions for support of multiple association groups are left for a future version of this document.

To create a new association group on a PCC, a PCE sends a PCUpd message which contains the LSP Object(e.g. PLSP-ID=100) and ASSOCIATION-GROUP TLV (Association Group ID=10) in the LSP object. Next, a PCE sends the another PCUpd message with another LSP Object(e.g. PLSP-ID=200) and ASSOCIATION-GROUP TLV(Association Group ID=10). As a result, the PCC and PCE both recognize that Association Group ID 10 represents PLSP-ID=100 and 200.

To remove a specific PLSP-ID from the association group, a PCE sends PCUpd message which contains the LSP Object(PLSP-ID=100) and ASSOCIATION-GROUP TLV (Association Group ID=0x0000). Then a PCC removes the PLSP-ID 100 from any association groups on the PCC.

To flush all association groups on a PCC, a PCE sends a PCUpd message which contains the LSP Object(PLSP-ID=0x0000) and ASSOCIATION-GROUP TLV(Association Group ID=0x0000). Then a PCC flushes all association groups.

## 5.2. DATA-CONTROL TLV in LSP Objects

This document defines DATA-CONTROL TLV in LSP Objects.

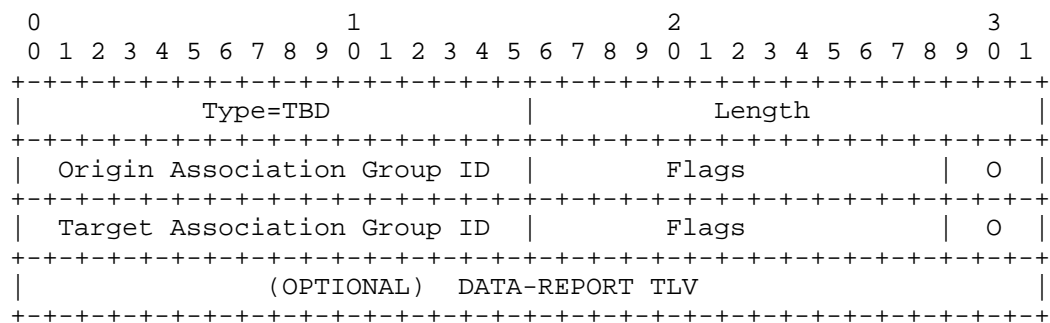


Figure 2: DATA-CONTROL TLV format

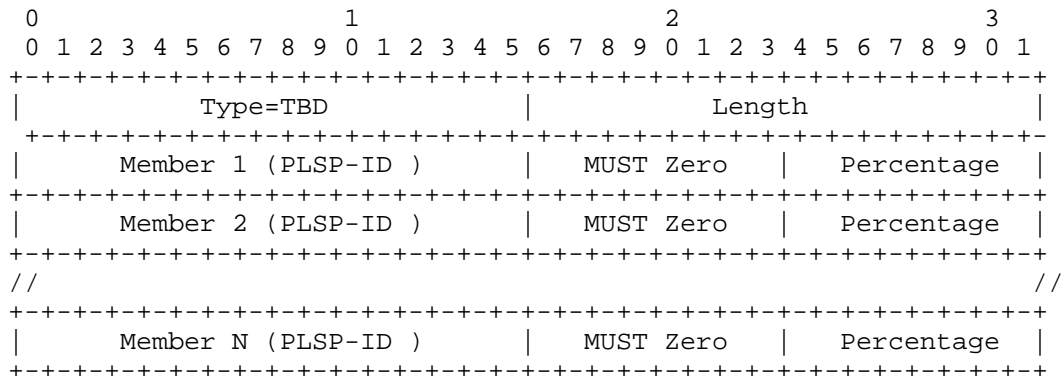


Figure 3: DATA-REPORT TLV format

## Flags and fields

- Origin Association Group ID: data traffic origin
- Target Association Group ID: for taking over whole data traffic from origin.
- O (Operational - 3 bit): This bit represents the requested operational status by a PCE. The meanings of the values are defined in [I-D.ietf-pce-stateful-pce].
- Member in DATA-REPORT TLV: This TLV is used in a PCRpt message and represents actual percentages of load balancing per respective PLSP-ID after load balancing execution. Member field fills PLSP-ID that is member of target association group.
- Percentage in DATA-REPORT TLV - 8 bit: This field specifies actual percentage of load balancing as an unsigned char.

An LSP Object in PCCUpd message MUST have DATA-CONTROL TLV when a PCE operates data switchover and balancing on a PCC. DATA-CONTROL TLV is sub-TLV of an LSP Object and is used in both PCUpd and PCRpt message.

An operation of data switchover/balancing is the action of transferring traffic from an origin association group to a target association group. A PCUpd message with reserved LSP Object (PLSP-ID=0x0000) and DATA-CONTROL TLV (a set of an origin and a target association group) MUST triggers data switchover/balancing execution.

A PCC replies to a PCE a PCRpt message as a notification of data ' switchover/balancing result. The PCRpt message MUST have reserved LSP Object(PLSP-ID=0x0000) and DATA-CONTROL TLV with DATA-REPORT inside.

## 6. Operation Examples

For easy understanding this section introduces typical operation examples of data switchover/balancing.

### 6.1. Data switchover operation (100:0 => 0:100)

A PCE instructs a PCC to switchover 100% traffic from association group ID 1 to association group ID 2. A PCE sends single PCUpd message containing the reserved LSP Objects with DATA-CONTROL TLV.

Expected PCUpd,PCRpt messages to create association group and to trigger data switchover follow.

PCE	PCC(Ingress)	Egress
[LSP Association for existing LSP]		
--PCUpd ----->		
LSP Obj: PLSP-ID=1		
+ ASSOC-G: Assoc-G-ID 10		
<---PCRpt -----		
LSP Obj: PLSP-ID=1		
+ ASSOC-G: Assoc-G-ID 10		
[LSP Creation]		
--PCInitiate ----->		
	--Path ----->	
<---PCRpt -----	<----- Resv--	Establish a new
LSP Obj: PLSP-ID=2		PCE-Initiated LSP
[LSP Association for PCE-Initiated LSP]		
--PCUpd ----->		
LSP Obj: PLSP-ID=2		
+ ASSOC-G: Assoc-G-ID 20		
<---PCRpt -----		
LSP Obj: PLSP-ID=2		
+ ASSOC-G: Assoc-G-ID 20		
[Switchover Execution]		
--PCUpd ----->		
LSP Obj: PLSP-ID=0x0000		
+ D-CTRL:	:	
Origin Assoc-G-ID 10(O=up)	:	
Target Assoc-G-ID 20(O=active)	:	
	)))))))))))	Switchover
	})))))))))	Execution
<---PCRpt-----	:	
LSP Obj: PLSP-ID=0x0000	:	
+ D-CTRL:	:	
Origin Assoc-G-ID 10(O=up)		
Target Assoc-G-ID 20(O=active)		
+ D-REPORT:		
PLSP-ID 2, 100%		

Figure 4: Switchover Operation Example

## 6.2. Load balancing operation (100:0 =&gt; 50:50)

The scenario is one where the starting state is a single LSP (of bandwidth 100 M) is carrying the traffic. To enable better bin-packing, the PCE may want to create two smaller LSPs instead, each of 50M, and load balance the traffic over them. To accomplish this, two association groups are used, the first (say association group ID 10) contains the LSP carrying the traffic, and the second (say association group ID 30) contains the two new smaller LSPs. Expected PCUpd, PCRpt messages to create association group and to trigger load-balance follow (The instantiation of the original LSP of bandwidth 100M and its association into group ID 10 is not shown)

PCE	PCC(Ingress)	Egress
[LSP Creation]		
--PCInitiate x2-----> BW: 50M	--Path x2-----> <-----Resv x2--	Establish two new PCE-Initiated LSP
<--PCRpt ----- LSP Obj: PLSP-ID=3		
<--PCRpt ----- LSP Obj: PLSP-ID=4		
[LSP Association for PCE-Initiated LSPs]		
--PCUpd -----> LSP Obj: PLSP-ID=3 + ASSOC-G: Assoc-G-ID 30		Create new Association Group for PCE-Initiated LSP
<--PCRpt ----- LSP Obj: PLSP-ID=3 + ASSOC-G: Assoc-G-ID 30		
--PCUpd -----> LSP Obj: PLSP-ID=4 + ASSOC-G: Assoc-G-ID 30		Add a new LSP to Association Group
<--PCRpt ----- LSP Obj: PLSP-ID=4 + ASSOC-G: Assoc-G-ID 30		

```
[Load Balancing Execution]
|--PCUpd----->
  LSP Obj: PLSP-ID=0x0000
  + D-CTRL:
    Origin Assoc-G-ID 10(O=up)
    Target Assoc-G-ID 30(O=active)
    )))
    )))
  <--PCRpt-----
    LSP Obj: PLSP-ID=0x0000
    + D-CTRL:
      Origin Assoc-G-ID 10(O=up)
      Target Assoc-G-ID 30(O=active)
      + D-REPORT:
        PLSP-ID 3, 50%
        PLSP-ID 4, 50%
```

Figure 5: Load-Balance Operation Example

## 7. IANA Considerations

## 7.1. PCEP TLV Indicators

This document defines the following new PCEP TLVs:

Value	Meaning	Reference
TBD	DATA-CONTROL	This document
TBD	DATA-REPORT	This document

## 7.2. PCEP Error Objects

This document defines new Error-Type and Error-Value for the following new error conditions:



Error-Type	Meaning
6	Mandatory Object missing Error-value=TBD: DATA-CONTROL TLV missing. Error-value=TBD: DATA-REPORT TLV missing.
19	Invalid operation Error-value=TBD: No association group existing. Error-value=TBD: No association group specified.

## 8. Security Considerations

TBD

## 9. Acknowledgments

Many thanks to Adrian Farrel for his ideas and suggestions.

## 10. References

### 10.1. Normative References

- [I-D.crabbe-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-crabbe-pce-pce-initiated-lsp-02 (work in progress), July 2013.
- [I-D.ietf-pce-stateful-pce]  
Crabbe, E., Medved, J., Minei, I., and R. Varga, "PCEP Extensions for Stateful PCE", draft-ietf-pce-stateful-pce-05 (work in progress), July 2013.
- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC4872] Lang, J., Rekhter, Y., and D. Papadimitriou, "RSVP-TE Extensions in Support of End-to-End Generalized Multi-Protocol Label Switching (GMPLS) Recovery", RFC 4872, May 2007.
- [RFC5440] Vasseur, JP. and JL. Le Roux, "Path Computation Element (PCE) Communication Protocol (PCEP)", RFC 5440,

March 2009.

## 10.2. Informative References

- [I-D.tanaka-pce-stateful-pce-mbb]  
Tanaka, Y. and Y. Kamite, "Make-Before-Break MPLS-TE LSP restoration and reoptimization procedure using Stateful PCE", draft-tanaka-pce-stateful-pce-mbb-00 (work in progress), February 2013.
- [RFC3209] Awduche, D., Berger, L., Gan, D., Li, T., Srinivasan, V., and G. Swallow, "RSVP-TE: Extensions to RSVP for LSP Tunnels", RFC 3209, December 2001.

## Authors' Addresses

Yosuke Tanaka  
NTT Communications Corporation  
Granpark Tower  
3-4-1 Shibaura, Minato-ku  
Tokyo 108-8118  
Japan

Email: yosuke.tanaka@ntt.com

Yuji Kamite  
NTT Communications Corporation  
Granpark Tower  
3-4-1 Shibaura, Minato-ku  
Tokyo 108-8118  
Japan

Email: y.kamite@ntt.com

Ina Minei  
Juniper Networks, Inc.  
1194 N. Mathilda Ave.  
Sunnyvale, CA 94089  
US

Email: ina@juniper.net



Network Working Group  
Internet-Draft  
Intended status: Standards Track  
Expires: January 16, 2014

Y. Tanaka  
Y. Kamite  
NTT Communications  
Jul 15, 2013

Make-Before-Break MPLS-TE LSP restoration and reoptimization procedure  
using Stateful PCE  
draft-tanaka-pce-stateful-pce-mbb-01

Abstract

Stateful PCE (Path Computation Element) and its corresponding protocol extensions provide a mechanism that enables PCE to do stateful control of MPLS Traffic Engineering Label Switched Paths (TE LSP). Stateful PCE supports manipulating the existing LSP's state and attributes (e.g., bandwidth and route) and also creating totally new LSPs in the network.

In the current MPLS TE network using RSVP-TE, LSPs are often controlled by "make-before-break (M-B-B)" signaling by headend for the purpose of LSP restoration and reoptimization. In most cases, it is an essential operation to reroute LSP traffic without any data disruption.

This document specifies the procedure of applying stateful PCE's control to make-before-break RSVP-TE signaling. In this document, two types of restoration/reoptimization procedures are defined, implicit mode and explicit mode. This document also specifies the usage and handling of stateful PCEP (PCE Communication Protocol) messages, expected behavior of PCC as RSVP-TE headend and several extensions of additional PCEP objects.

Status of this Memo

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

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

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

This Internet-Draft will expire on January 16, 2014.

#### Copyright Notice

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

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

## Table of Contents

1. Introduction . . . . .	4
2. Conventions used in this document . . . . .	4
3. Terminology . . . . .	4
4. Motivation . . . . .	5
5. Make-Before-Break LSP procedures . . . . .	5
5.1. Implicit Make-Before-Break Mode . . . . .	6
5.2. Explicit Make-Before-Break Mode . . . . .	8
5.2.1. Establish new Trial LSP . . . . .	9
5.2.2. Switchover Data Traffic triggered by a PCUpd message . . . . .	10
5.2.3. Tear Down old LSP . . . . .	12
6. Objects and TLV Formats . . . . .	12
6.1. Trial LSP TLV in LSP Objects . . . . .	12
6.2. DATA-CONTROL TLV in LSP Objects . . . . .	12
7. IANA Considerations . . . . .	13
7.1. PCEP TLV Indicators . . . . .	13
7.2. PCEP Error Objects . . . . .	14
8. Security Considerations . . . . .	14
9. Acknowledgments . . . . .	14
10. References . . . . .	14
10.1. Normative References . . . . .	14
10.2. Informative References . . . . .	15
Authors' Addresses . . . . .	15

## 1. Introduction

[I-D.ietf-pce-stateful-pce] describes the stateful Path Computation Elements(PCE). Stateful PCE defines the extensions to PCEP to enable stateful control of LSPs between and across PCEP sessions, and it also describes mechanisms to effect LSP state synchronization between PCCs and PCEs, and PCE control of timing and sequence of path computations within and across PCEP sessions.

Today, however, there is no detailed procedure specified as to how to restore and reoptimize one particular MPLS-TE LSP using stateful PCE. In today's MPLS RSVP-TE mechanism, make-before-break (M-B-B) is a widely common scheme supported by headend LER in order to assure no traffic disruption during restoration and reoptimization. Hence it is naturally desirable for stateful PCE to control M-B-B based signaling and forwarding process.

This document specifies the definite procedures of applying stateful PCE's control to M-B-B method. In this document, two types of restoration/reoptimization procedures are defined, Implicit mode and explicit mode. This document also specifies the usage and handling of stateful PCEP (PCE Communication Protocol) messages, expected behavior of PCC as RSVP-TE headend and several extensions of additional objects.

## 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].

## 3. Terminology

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

This document uses the following terms defined in [RFC3209]: make-before-break.

This document uses the following terms defined in [RFC4426] and [RFC4427]: recovery, protection, restoration.

According to their definition the term "recovery" is generically used to denote both protection and restoration; the specific terms "protection" and "restoration" are used only when differentiation is required. The subtle distinction between protection and restoration

is made based on the resource allocation done during the recovery period. Hence the protection allocates LSP resource in advance of a failure, while the restoration allocates LSP after a failure occurs.

#### 4. Motivation

As for current MPLS mechanism, make-before-break(M-B-B) concept is outlined in [RFC3209], which allows adaptive and smooth RSVP-TE LSP rerouting that does not disrupt traffic or adversely impact network operations while rerouting is in progress. M-B-B is applicable for reoptimizing LSP's route and resources for several use cases, for example, to adopt better path for reversion after failure, to change traversing node/links for planned maintenance, to change bandwidth of LSPs. M-B-B is also used for global restoration scenario in case of failure, which is effective if operators do not want to reserve both working and standby LSPs' bandwidth in advance. In real deployment, it can also be operated with local protection scheme FRR (Fast ReRoute).

Since M-B-B operational scheme is universally common in MPLS network today, it is naturally much desirable to utilize it under the architecture of stateful PCE.

The basic procedure of the Make-Before-Break method is outlined as follows:

1. Establish a new LSP
2. Transfer data traffic from old LSP onto the new LSP
3. Tear down the old LSP

In M-B-B, it is an important behavior that headend node handles the sequence of data traffic switchover. The headend is able to "make" one or more new LSPs for a particular Tunnel (i.e., it is allowed to signal multiple RSVP sessions with different LSP-IDs that share a common Tunnel IDs), and the headend will switch the traffic upon only one (or some) of those LSPs. In some use cases about stateful PCE, it is expected that operators can watch and control when the data is switched over and which LSPs are used. Therefore, this document covers such a procedure and related message extensions.

#### 5. Make-Before-Break LSP procedures

There are possibly two modes introduced for Make-Before-Break procedure under stateful PCE. The first one is "implicit M-B-B mode", where the operation is triggered by a PC Update Request(PCUpd) message from a PCE, and a PCC handles whole Make-Before-Break steps



(signaling and transferring data traffic) for itself. This mode utilizes the existing messages as defined in [I-D.ietf-pce-stateful-pce] .

The second one is "explicit M-B-B mode", where the operation is triggered by a PCUpd message with TRIAL LSP TLV, which is defined in Section 6.1. A PCE also controls timing and sequence of each granular step that a PCC takes. This procedure additionally uses a new extended TLV that is defined in Section 6.2

Both types of procedure require at least two LSPs residing in a single MPLS-TE tunnel, working LSP and trial LSPs. An ingress node is currently transporting data traffic on the working LSP, and then it establishes one or more trial LSPs. As per [RFC3209] Section 2.5. "LSP ID" of a restoration LSP, which is newly signaled, differs from that of a working LSP. In this document, LSP ID of a working LSP describes "old" and that of a trial LSP describes "new" as a simple example.

Implicit mode has high affinity with most existing MPLS edge node implementations which perform entire steps of M-B-B automatically at once. This mode is particularly applicable for migration scenario for the existing deployment where service providers want their recovery operation be delegated to centralized PCE.

Explicit mode is much more flexible than Implicit mode since it allows PCEs to manage each LSP step-by-step. Explicit mode is applicable to several new use cases that require split control of signaling and data switchover. For example, if end-to-end data path is created by connecting multiple individual LSPs across different segments (e.g., LSP stitching), in reoptimization scenario, data flowing cannot be started unless all signaling of all LSPs is completed. Similarly, there is a case under Software Defined Network (SDN) applications, where MPLS domain is connected to other non-MPLS domains, and the end-to-end data switchover timing should be carefully coordinated with various different methods of path/flow setup in each domain.

PCC and PCE can distinguish which mode, implicit mode or explicit mode, is to be performed by checking the type of PCEP messages that are exchanged. The implementation MAY support both modes, but for each restoration/reoptimization operation, either one of them SHOULD be exclusively selected.

#### 5.1. Implicit Make-Before-Break Mode

This specifies the detailed procedure of M-B-B LSP restoration and reoptimization using existing messages which are defined in

[I-D.ietf-pce-stateful-pce] . This procedure is based on the current existing messages/TLVs and no extended TLV is used. Once a PCC receives PCUpd message from a PCE, the PCC automatically executes the implicit M-B-B procedure as described in [I-D.ietf-pce-stateful-pce] Section 6.2.

First, A PCUpd message is sent from a PCE to trigger M-B-B procedure. Once a PCC received the PCUpd message, the PCC starts signaling a new restoration LSP and it sends back to the PCE a PCRpt message with LSP-IDENTIFIERS TLV in the LSP Object.

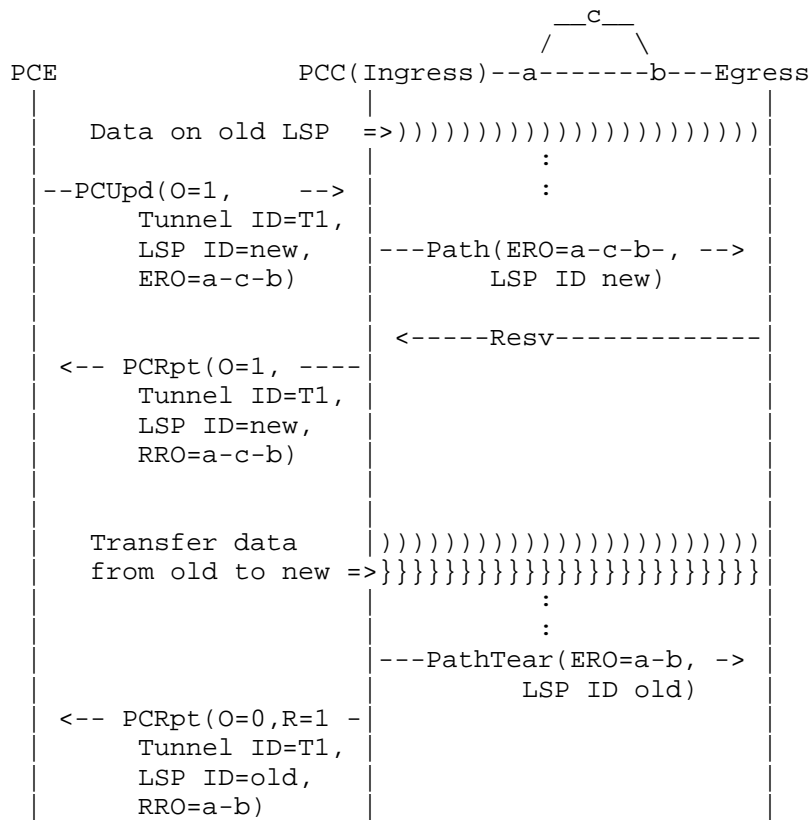
Second, once a restoration LSP is successfully established, a PCC transfers data traffic from working LSP to restoration LSP. If the restoration LSP failed in setup, the PCC notifies the PCE the result in a PCRpt message and it MAY wait for a next instruction from the PCE.

Finally, when a PCC successfully transferred data traffic to restoration LSP, the PCC tears down the (previous) working LSP by RSVP-TE signaling, then the PCC MUST send a PCRpt message. That PCRpt message MUST carry a LSP Object with LSP-IDENTIFIERS TLV which indicates the value of RSVP-TE signaling the PCC has just torn down.

Following Figure 1 illustrates the example of implicit M-B-B procedure, in following conditions.

working LSP : ERO=a-b, Tunnel ID=T1, LSP ID=old

restoration LSP : ERO=a-c-b, Tunnel ID=T1, LSP ID=new



O flag = Operational flag in LSP object.  
 R flag = Remove flag in LSP object.

Figure 1: Implicit Make-Before-Break Procedure

## 5.2. Explicit Make-Before-Break Mode

Comparing to the implicit M-B-B mode, explicit M-B-B mode allows a PCE to control timing and sequence of subsequent make-before-break steps as follows.

First, the PCE initiates PCC's signaling of a new LSP by sending a LSP Update Request(PCUpd) message with TRIAL-LSP TLV that are defined in this document. Second, the PCE instructs the PCC to transfer data traffic from old LSP to new LSP by sending a PCUpd message with TRIAL-LSP TLV and DATA-CONTROL TLV that are defined in [I-D.tanaka-pce-stateful-pce-data-ctrl]. Third, the PCE MAY instruct the PCC to

tear down the old LSP by sending a PCUpd message indicating LSP removal.

The following subsections specify each Make-Before-Break steps in detail.

#### 5.2.1. Establish new Trial LSP

As a first step of M-B-B procedure, a PCC establishes a new LSP for restoration once PCC receives a PCUpd message with TRIAL-LSP TLV from a PCE. We call this newly established LSPs for restoration "trial LSP". A trial LSP is signaled the same RSVP-TE Tunnel ID but different LSP ID from active working LSP, and both the active working LSP and new trial LSPs MUST be signaled with Shared Explicit style as describes in [RFC3209]. TRIAL-LSP TLV triggers explicit mode M-B-B. A PCE do not have to assign RSVP-TE LSP ID for trial LSP signaling, however it MAY specify RSVP-TE LSP ID that the PCC is going to establish.

When a new trial LSP was signaled successfully, the PCC sends a PCRpt message toward the PCE to notify the result. The PCRpt message from the PCC MUST have the LSP object with LSP-IDENTIFIERS TLV that indicates RSVP-TE Tunnel ID and LSP ID the PCC has just established.

If a new trial LSP failed to be established by some reason of RSVP-TE signaling, the PCC MUST send a PCRpt message carrying LSP-IDENTIFIERS TLV and RSVP-ERROR-SPEC TLV as defined in [I-D.ietf-pce-stateful-pce] Section 7.3.4. to the PCE.

A PCC SHOULD accept multiple PCUpd messages with TRIAL-LSP TLV in a LSP Object. And a PCC SHOULD establish as many trial lsps as the number of PCUpd messages it receives.

Figure 2 illustrates a example, working LSP(PLSP-ID P1,Tunnel ID T1, LSP-ID old, ERO Ingress-a-b-Egress), trial LSP(Tunnel ID T1, LSP-ID new, ERO Ingress-a-c-b-Egress).

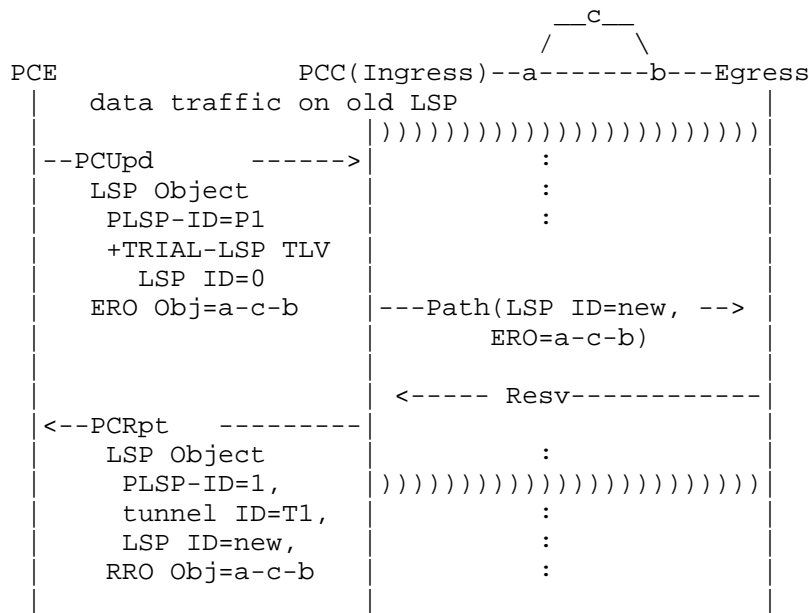


Figure 2: Establish new LSP

### 5.2.2. Switchover Data Traffic triggered by a PCUpd message

As a second step, PCC(Ingress) transfers data traffic from a working LSP to a trial LSP. To specify desired LSP for transferring data traffic, a PCUpd message from a PCE MUST have a TRIAL-LSP TLV and a DATA-CONTROL TLV in a LSP Object.

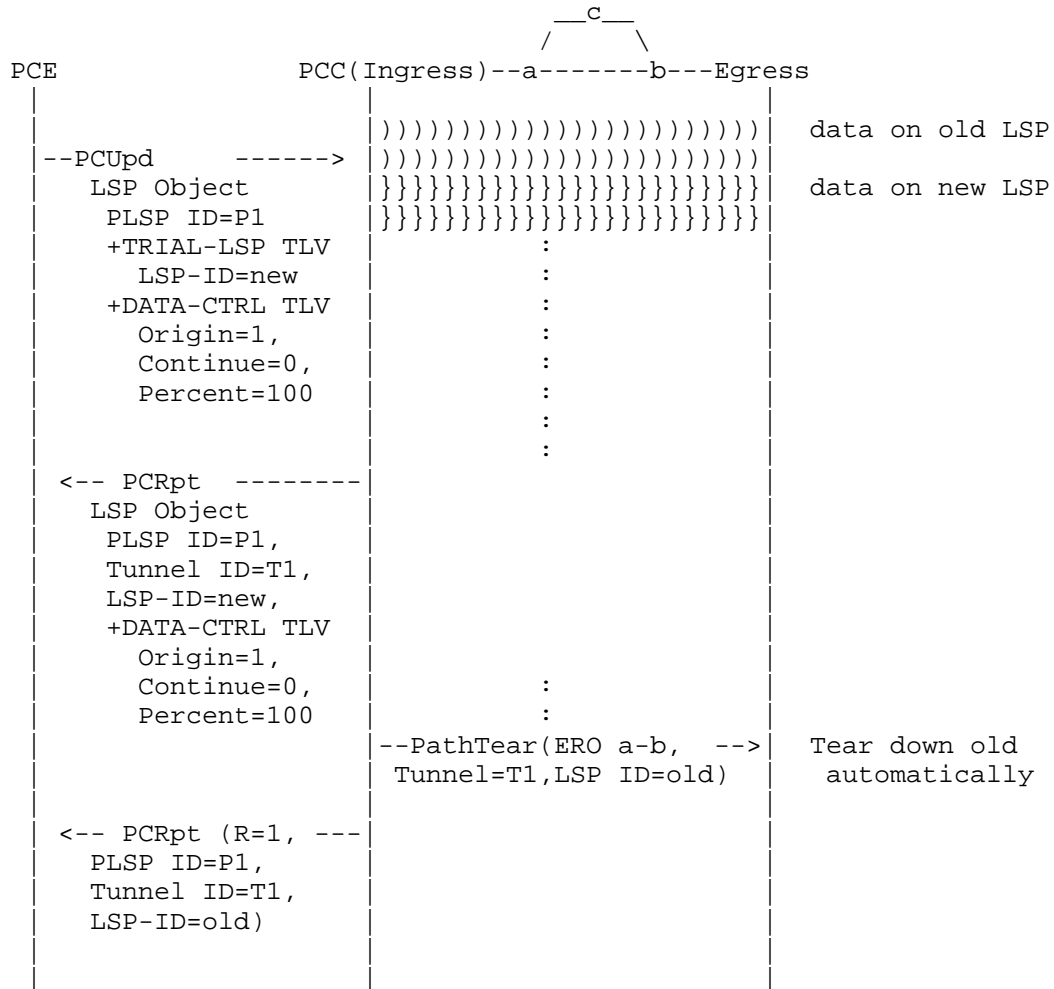
A TRIAL-LSP TLV specifies desired RSVP-TE LSP-ID a PCC starts using for transferring data traffic. And a DATA-CONTROL TLV triggers data traffic switch over. Both TLVs MUST be assembled into a single LSP Object.

Once the PCC receives the PCUpd message with TRIAL-LSP TLV and DATA-CONTROL TLV in the LSP Object, the PCC MUST start transfer data traffic to new trial LSP immediately. (See Figure 3)

In DATA-CONTROL TLV, Origin(O) bit, which represents traffic origin, SHOULD set to 1, Continue(C) bit SHOULD set to 0, and Percentage(P) bit SHOULD set to 100% in order to perform whole data traffic switchover.

If the TRIAL-LSP TLV in the PCUpd message specifies invalid LSP,

PCErr MUST be sent out from the PCC to the PCE. The error message with Error-Type-19 (Invalid Operation) and Error-Value[TBD](See Section 7.2.



R flag = Remove flag in LSP object.

Figure 3: Transfer data traffic from old LSP to new LSP

### 5.2.3. Tear Down old LSP

As a final step of Make-Before-Break procedure, the PCC tears down the working LSP and the other trial lsps which the data traffic is no longer used.

The PCC SHOULD tear down the old working LSP and other trial LSPs immediately once the data traffic succesfully switched over (See Figure 3). In OPTIONAL, a PCC tears down old lsp separately.

## 6. Objects and TLV Formats

### 6.1. Trial LSP TLV in LSP Objects

This document defines a new TLV named TRIAL-LSP TLV.

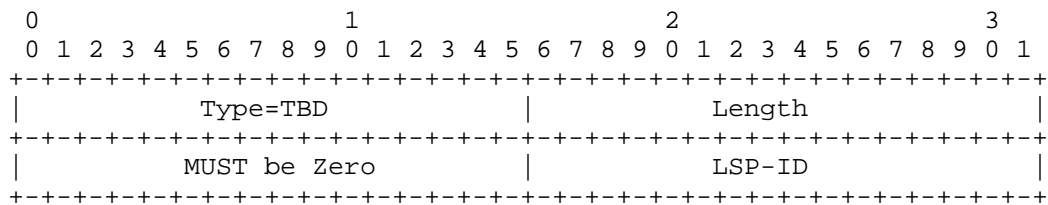


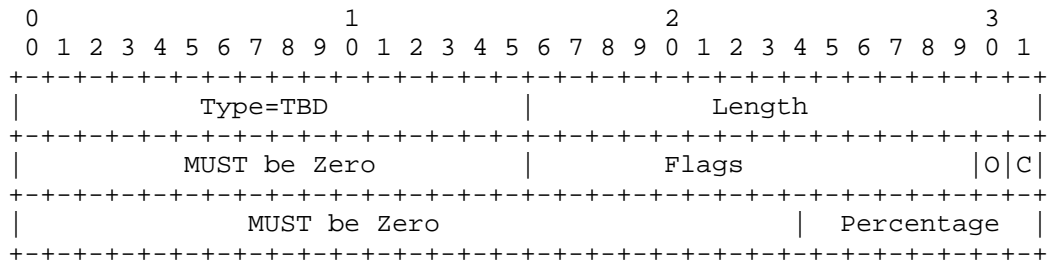
Figure 4: TRIAL-LSP TLV format

TRIAL-LSP TLV is sub-TLV of the LSP Object and is used in a PCUpd message especially to perform explicit mode M-B-B. A PCC signals a trial LSP once it receives a PCUpd in which LSP object has a TRIAL-LSP TLV(LSP-ID=0). It MUST set RSVP LSP-ID in LSP-ID field of TRIAL-LSP TLV in order to notify a PCC of desired trial LSP to be carried data traffic.

**LSP-ID:** This field fills the same value of RSVP-TE LSP-ID that is used in signaling. LSP-ID MUST be zero in a PCUpd message when a PCE requests a PCC to signal new trial LSP. LSP-ID MUST be non-zero when a PCE sends a PCUpd message to trigger traffic switchover execution.

### 6.2. DATA-CONTROL TLV in LSP Objects

DATA-CONTROL TLV in LSP Objects follows for easy reference of [I-D.tanaka-pce-stateful-pce-data-ctrl].



## DATA-CONTROL TLV format

## Flags and fields

- O (traffic Origin - 1 bit): "traffic Origin(0) = 1" indicates this is an active LSP (i.e., carrying traffic now) whose traffic is to be switched over or to be load-balanced. A PCE uses this bit to specify traffic origin that it wants to manipulate. On the other hand, A PCC uses this bit in PCRpt message to notify a PCE that switching traffic succeeded and carrying data traffic.
- C (Continue - 1 bit): If this flag set to 1, it indicates the next LSP Object encoded in the PCUpd has also DATA-CONTROL TLV. If this flag set to 0, it indicates no more LSP Object continues and load balancing calculation is completed, and then the PCC MUST perform switching traffic or load-balancing.
- Percentage - 8 bit [0B11111111 is reserved]: This field specifies ratio of switching traffic as an unsigned char. The sum of this field across subsequent LSP Object has to be hundred percent. The value must be less than or equal to 100% (0B01100100) (e.g., If you want to set 50%, this field should be set to 0B00110010). If no traffic goes through the corresponding LSP, this field should be set to 0%. 0% LSP MUST be deleted immediately after switchover. The special value 0B11111111 indicates traffic 0%, but the LSP MUST remain after switchover.

## 7. IANA Considerations

## 7.1. PCEP TLV Indicators

This document defines the following new PCEP TLVs:

Value	Meaning	Reference
TBD	DATA-CONTROL	This document



## 7.2. PCEP Error Objects

This document defines new Error-Type and Error-Value for the following new error conditions:

Error-Type	Meaning
6	Mandatory Object missing Error-value=TBD: LSP Identifiers TLV missing
19	Invalid operation Error-value=TBD: Percentage is not hundred. for explicit mode Error-value=TBD: Specified LSP-ID is not existing. for explicit mode Error-value=TBD: Specified LSP-ID is not operational. for explicit mode

## 8. Security Considerations

TBD

## 9. Acknowledgments

Many thanks to Ina Minei, Adrian Farrel, Yimin Shen, Xian Zhang and their develop team for their ideas and feedback in documentation.

## 10. References

### 10.1. Normative References

- [I-D.ietf-pce-stateful-pce]  
Crabbe, E., Medved, J., Minei, I., and R. Varga, "PCEP Extensions for Stateful PCE",  
draft-ietf-pce-stateful-pce-05 (work in progress),  
July 2013.
- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC5440] Vasseur, JP. and JL. Le Roux, "Path Computation Element (PCE) Communication Protocol (PCEP)", RFC 5440, March 2009.

## 10.2. Informative References

- [RFC3209] Awduche, D., Berger, L., Gan, D., Li, T., Srinivasan, V., and G. Swallow, "RSVP-TE: Extensions to RSVP for LSP Tunnels", RFC 3209, December 2001.
- [RFC4426] Lang, J., Rajagopalan, B., and D. Papadimitriou, "Generalized Multi-Protocol Label Switching (GMPLS) Recovery Functional Specification", RFC 4426, March 2006.
- [RFC4427] Mannie, E. and D. Papadimitriou, "Recovery (Protection and Restoration) Terminology for Generalized Multi-Protocol Label Switching (GMPLS)", RFC 4427, March 2006.

## Authors' Addresses

Yosuke Tanaka  
NTT Communications Corporation  
Granpark Tower  
3-4-1 Shibaura, Minato-ku  
Tokyo 108-8118  
Japan

Email: yosuke.tanaka@ntt.com

Yuji Kamite  
NTT Communications Corporation  
Granpark Tower  
3-4-1 Shibaura, Minato-ku  
Tokyo 108-8118  
Japan

Email: y.kamite@ntt.com



PCE Working Group  
Internet-Draft  
Intended status: Standards Track  
Expires: January 16, 2014

Q. Wu  
D. Dhody  
Huawei  
July 15, 2013

Path Computation Element (PCE) Discovery using Domain Name System(DNS)  
draft-wu-pce-dns-pce-discovery-01

Abstract

Discovery of the Path Computation Element (PCE) within an IGP area or domain is possible using OSPF [RFC5088] and IS-IS [RFC5089]. However, in some deployment scenarios PCEs may not wish, or be able, to participate within the IGP process, therefore it would be beneficial for the Path Computation Client (PCC) (or other PCEs) to discover PCEs via an alternative mechanism to those proposed in [RFC5088] and [RFC5089].

This document specifies the requirements, use cases, procedures and extensions to support discovery via DNS for 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."

This Internet-Draft will expire on January 16, 2014.

Copyright Notice

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

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

carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

## Table of Contents

1. Introduction . . . . .	3
1.1. Requirements . . . . .	3
2. Conventions used in this document . . . . .	4
3. Motivation . . . . .	5
3.1. Load Sharing of Path Computation Requests . . . . .	5
3.2. Network Address Translation Gateway . . . . .	5
3.3. Multiple-Provider Domains . . . . .	5
3.4. Multiple PCE Servers . . . . .	6
3.5. End to End Path Computation . . . . .	6
4. Discovering a Path Computation Element . . . . .	7
4.1. Determining the PCE Service and transport protocol . . . . .	8
4.2. Determining the IP Address of the PCE . . . . .	8
4.3. Determining path computation scope,the PCE domains and Neighbor PCE domains . . . . .	9
5. IANA Considerations . . . . .	11
6. Security Considerations . . . . .	12
7. Acknowledges . . . . .	13
8. References . . . . .	14
8.1. Normative References . . . . .	14
8.2. Informative References . . . . .	15
Authors' Addresses . . . . .	16

## 1. Introduction

The Path Computation Element Communication Protocol (PCEP) is a transaction-based protocol carried over TCP[RFC4655]. In order to be able to direct path computation requests to the Path Computation Element (PCE), a Path Computation Client (PCC) (or other PCEs) needs to know the location and capability of a PCE.

In a network where an IGP is used and where the PCE participates in the IGP, discovery mechanisms exist for PCC (or PCE) to learn the identity and capability of each PCE. [RFC5088] defines a PCE Discovery (PCED) TLV carried in an OSPF Router LSA. Similarly, [RFC5089] defines the PCED sub-TLV for use in PCE Discovery using IS-IS. Scope of the advertisement is limited to IGP area/level or Autonomous System (AS).

However in certain scenarios not all PCEs will participate in the IGP instance, section 3 (Motivation) outlines a number of use cases. In these cases, current PCE Discovery mechanisms are therefore not appropriate and another PCE discovery function would be required.

### 1.1. Requirements

As described in [RFC4674], the PCE Discovery information should at least be composed of:

- o The PCE location: an IPv4 and/or IPv6 address that is used to reach the PCE. It is RECOMMENDED to use an address that is always reachable if there is any connectivity to the PCE;
- o The PCE path computation scope (i.e., intra-layer, inter-area, inter-AS, or inter-layer);
- o The set of one or more PCE-Domain(s) into which the PCE has visibility and for which the PCE can compute paths;
- o The set of zero, one, or more neighbor PCE-Domain(s) toward which the PCE can compute paths;

that allows PCCs to select appropriate PCEs:

This document specifies an extension to DNS for the above PCE information discovery, which complements the existing discovery mechanism.

## 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].

### 3. Motivation

This section discusses in more detail the motivation and use cases for an alternative DNS based PCE discovery mechanism.

#### 3.1. Load Sharing of Path Computation Requests

Multiple PCE servers can be present in a single network domain for redundancy. However load balance decision is made by PCC, it doesn't enable real load balance across the PCE servers if PCC still tries PCE one by one and PCE doesn't indicate the load status to the PCC.

Inherent DNS based load balancing may be used for inbound load balancing and implemented at the application level in both servers and clients. Multiple host IP addresses are configured in DNS for a single host server name. Also DNS is query-response based mechanism and capable of automatically detecting and reacting to errors. These allow you to provide load balancing across two separate Systems and facilitate PCE system failover and recovery.

Comparing with advertisement based PCE discovery [RFC5088][RFC5089], it can mitigate flooding issue (see section 3.2 of [RFC5088]) and avoid unwanted traffic and reduce a large amount of unnecessary advertisement, especially when PCE information needs frequent changes.

#### 3.2. Network Address Translation Gateway

PCEP uses TCP as the transport [RFC5440]. To secure TCP connection that underly PCEP sessions, TLS can be used besides using TCP-MD5. When NAT gateway is in place, a TCP or TCP/TLS connection can be opened by ICE for the purpose of connectivity checks. However the TCP connection cannot be established in cases where one of the agents is behind a NAT with connection-dependent filtering properties [RFC5382]. Therefore IGP discovery is limited within an IGP domain and cannot be used in this case.

#### 3.3. Multiple-Provider Domains

Backward recursive path computation (BRPC) [RFC5441] MAY be used by cooperating PCEs to compute inter-domain path. In which case these cooperating PCEs should known to other PCEs. In case of inter-AS where the PCE do not participate in a common IGP, the existing IGP discovery mechanism cannot be used to discover inter-AS PCE.

Also in the case of multiple ASes within different service provider networks, the H-PCE [RFC6805] architecture does not require disclosure of internals of a child domain to the parent PCE. It may



be necessary for a third party to manage the parent PCEs according to commercial and policy agreements from each of the participating service providers.

[RFC6805] specifies that a child PCE must be configured with the address of its parent PCE in order for it to interact with its parent PCE. However handling changes in parent PCE identities and coping with failure events would be an issue for a configured system.

There is no scope for parent PCEs to advertise their presence, however there is potential for directory systems (such as DNS [RFC4848] as used in the ALTO discovery function [I-D.ietf-alto-server-discovery]).

### 3.4. Multiple PCE Servers

In some cases, each network domain may have multiple PCE server, only one main PCE sever is responsible for Establish topology database by participating in OSPF/ISIS routing protocol, the other PCE server gains knowledge of Topology information either from TED maintained by the main PCE server or some management system(e.g.,NMS/OSS). In such cases, it is desirable to use DNS based mechanism to discover PCE.

### 3.5. End to End Path Computation

To compute end to end paths across domains, PCE has the following limitations:

- o Within a single area, the PCE can not offers enhanced computational power for end to end path computation,e.g., coordination of computation across the whole area.
- o A single router participating in IGP area lacks visibility of complete topology with its own TED.

Per domain path computation mechanism[RFC5152]can be used to compute end to end path, however it may lead to sub-optimal paths or result in no end to end path to be found when the PCE only has visibility into the IGP area it serves. This issue can be resolved when one powerful PCE is responsible for multiple areas,i.e., PCE sits in one area it serves and also can get access to topology information provided by PCE server in other IGP area using BGP. In such case, it will be desirable to use DNS based mechanism to discover those PCE that has visibility to multiple areas.

#### 4. Discovering a Path Computation Element

The Dynamic Delegation Discovery System (DDDS) [RFC3401] is used to implement lazy binding of strings to data, in order to support dynamically configured delegation systems. The DDDS functions by mapping some unique string to data stored within a DDDS database by iteratively applying string transformation rules until a terminal condition is reached. When DDDS uses DNS as a distributed database of rules, these rules are encoded using the Naming Authority Pointer (NAPTR) Resource Record (RR). One of these rules is the First Well Known Rule, which says where the process starts.

In current specifications, the First Well Known Rule in a DDDS application [RFC3403] is assumed to be fixed, i.e., the domain in the tree where the lookups are to be routed to, is known. This document proposes the input to the First Well Known Rule to be dynamic, based on the search path the resolver discovers or is configured to use.

The search path of the resolver can either be pre-configured, or discovered using DHCP.

When the PCC or other PCEs needs to discover PCEs in the domain into which the PCEP speaker has visibility (e.g., local domain), the input to the First Well Known Rule MUST be the domain the PCC knows, which is assumed to be pre-configured in the PCC or discovered using DHCP.

When the PCC needs to discover PCE in the other domain (e.g., AS, Parent PCE in the parent domain) into which the PCC has no visibility, it SHOULD know the domain name of that domain and use DHCP to discover IP address of the PCE in that domain that provides path computation service along with some PCE location information useful to a PCC for PCE selection, and contact it directly. In some instances, the discovery may result in a per protocol/application list of domain names that are then used as starting points for the subsequent NAPTR lookups. If neither the IP address or PCE location information can be discovered with the above procedure, the PCC MAY request a domain search list, as described in [RFC3397] and [RFC3646], and use it as input to the DDDS application.

When the PCC does not find valid domain names using the procedures above, it MUST stop the attempt to discover any PCE.

The dynamic rule described above SHOULD NOT be used for discovering services other than Path computation services described in this document, unless stated otherwise by a future specification.

The procedures defined here result in an IP address, PCE domain, neighboring PCE domain and PCE Computation Scope where the PCC can

contact the PCE that hosts the service the PCC is looking for.

#### 4.1. Determining the PCE Service and transport protocol

The PCC should know the service identifier for the Path Computation Discovery service. The service identifier for the Path Computation Discovery service is defined as "PCED", The PCE supporting "PCED" service MUST support only TCP as transport, as described in [RFC5440].

The services relevant for the task of transport protocol selection are those with NAPTR service fields with values "ID+M2X", where ID is the service identifier defined in the previous section, and X is a letter that corresponds to a transport protocol supported by the domain. This specification only defines M2T for TCP. This document also establishes an IANA registry for mappings of NAPTR service name to transport protocol.

These NAPTR [RFC3403] records provide a mapping from a domain to the SRV [RFC2782] record for contacting a PCE with the specific transport protocol in the NAPTR services field. The resource record MUST contain an empty regular expression and a replacement value, which indicates the domain name where the SRV record for that particular transport protocol can be found. As per [RFC3403], the client discards any records whose services fields are not applicable.

The PCC MUST discard any service fields that identify a resolution service whose value is not "M2T", for values of T that indicate TCP transport protocols supported by the client. The NAPTR processing as described in RFC 3403 will result in the discovery of the most preferred transport protocol of the PCE that is supported by the client, as well as an SRV record for the PCE.

#### 4.2. Determining the IP Address of the PCE

As an example, consider a client that wishes to find "PCED" service in the example.com domain. The client performs a NAPTR query for that domain, and the following NAPTR records are returned:

Order	Pref	Flags	Service	Regexp	Replacement
1	IN	NAPTR	50	50	"s" "PCED" ""
					_PCED._tcp.example.com
2	IN	NAPTR	90	50	"s" "PCED+M2T" ""
					_PCED._tcp.example.com

This indicates that the domain does have a PCE providing Path Computation services over TCP, in that order of preference. Since the client only supports TCP, TCP will be used, targeted to a host

determined by an SRV lookup of `_PCED._tcp.example.com`. That lookup would return:

```
;; Priority Weight Port Target
IN SRV 0 1 XXXX server1.example.com
IN SRV 0 2 XXXX server2.example.com
```

where XXXX represents the port number at which the service is reachable.

Note that the regexp field in the NAPTR example above is empty. The regexp field MUST NOT be used when discovering path computation services, as its usage can be complex and error prone. Also, the discovery of the path computation service does not require the flexibility provided by this field over a static target present in the TARGET field.

If the client is already configured with the information about which transport protocol is used for a path computation service in a particular domain, it can directly perform an SRV query for that specific transport using the service identifier of the path computation Service. For example, if the client knows that it should be using TCP for path computation service, it can perform a SRV query directly for `_PCED._tcp.example.com`.

Once the server providing the desired service and the transport protocol has been determined, the next step is to determine the IP address.

According to the specification of SRV RRs in [RFC2782], the TARGET field is a fully qualified domain name (FQDN) that MUST have one or more address records; the FQDN must not be an alias, i.e., there MUST NOT be a CNAME or DNAME RR at this name. Unless the SRV DNS query already has reported a sufficient number of these address records in the Additional Data section of the DNS response (as recommended by [RFC2782]), the PCC needs to perform A and/or AAAA record lookup(s) of the domain name, as appropriate. The result will be a list of IP addresses, each of which can be contacted using the transport protocol determined previously.

#### 4.3. Determining path computation scope, the PCE domains and Neighbor PCE domains

DNS servers MAY use DNS TXT record and add new RRsets to the additional information section that are relevant to the answer and have the same authenticity as the data (the IP Address of the PCE) in the answer section. RRsets include path computation scope, the PCE domains and Neighbor PCE domains associated with the PCE. the PCC MAY

inspect those Additional Information section and be capable of handling responses from nameservers that never fill in the Additional Information part of a response.

## 5. IANA Considerations

The usage of NAPTR records described here requires well-known values for the service fields for the transport supported by Path Computation Services. The table of mappings from service field values to transport protocols is to be maintained by IANA.

The registration in the RFC MUST include the following information:

Service Field: The service field being registered.

Protocol: The specific transport protocol associated with that service field. This MUST include the name and acronym for the protocol, along with reference to a document that describes the transport protocol.

Name and Contact Information: The name, address, email address, and telephone number for the person performing the registration.

The following values have been placed into the registry:

Service Fields	Protocol
PCED+M2T	TCP

New Service Fields are to be added via Standards Action as defined in [RFC5226].

IANA is also requested to register PCED as service name in the Protocol and Service Names registry.

## 6. Security Considerations

It is believed that this proposed DNS extension introduces no new security considerations (i.e., A list of known threats to services using DNS) beyond those described in [RFC3833]. For most of those identified threats, the DNS Security Extensions [RFC4033] does provide protection. It is therefore recommended to consider the usage of DNSSEC [RFC4033] and the aspects of DNSSEC Operational Practices [RFC4641] when deploying Path Computation Services.

In deployments where DNSSEC usage is not feasible, measures should be taken to protect against forged DNS responses and cache poisoning as much as possible. Efforts in this direction are documented in [RFC5452].

Where inputs to the procedure described in this document are fed via DHCP, DHCP vulnerabilities can also cause issues. For instance, the inability to authenticate DHCP discovery results may lead to the Path Computation service results also being incorrect, even if the DNS process was secured.

## 7. Acknowledges

The author would like to thank Claire Bi, Ning Kong and Liang Xia for their review and comments that help improvement to this document.



## 8. References

### 8.1. Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", March 1997.
- [RFC2782] Gulbrandsen, A., "A DNS RR for specifying the location of services (DNS SRV)", RFC 2782, February 2000.
- [RFC3397] Aboba, B., "Dynamic Host Configuration Protocol (DHCP) Domain Search Option", RFC 3397, November 2002.
- [RFC3403] Mealling, M., "Dynamic Delegation Discovery System (DDDS) Part Three: The Domain Name System (DNS) Database", RFC 3403, October 2002.
- [RFC3646] Droms, R., "DNS Configuration options for Dynamic Host Configuration Protocol for IPv6 (DHCPv6)", RFC 3646, December 2003.
- [RFC4033] Arends, R., "DNS Security Introduction and Requirements", RFC 4033, March 2005.
- [RFC4641] Kolkman, O., "DNSSEC Operational Practices", RFC 4641, September 2006.
- [RFC4674] Droms, R., "Requirements for Path Computation Element (PCE) Discovery", RFC 4674, December 2003.
- [RFC4848] Daigle, D., "Domain-Based Application Service Location Using URIs and the Dynamic Delegation Discovery Service (DDDS)", RFC 4848, April 2007.
- [RFC5226] Narten, T., "Guidelines for Writing an IANA Considerations Section in RFCs", RFC 5226, May 2008.
- [RFC5440] Le Roux, J.L., "Path Computation Element (PCE) Communication Protocol (PCEP)", RFC 5440, April 2007.
- [RFC6805] King, D. and A. Farrel, "The Application of the Path Computation Element Architecture to the Determination of a Sequence of Domains in MPLS and GMPLS", RFC 6805, November 2012.

## 8.2. Informative References

- [ALTO] Kiesel, S., "ALTO Server Discovery",  
ID draft-ietf-alto-server-discovery-08, March 2013.
- [RFC3401] Mealling, M., "Dynamic Delegation Discovery System (DDDS)  
Part One: The Comprehensive DDDS", RFC 3401, October 2002.
- [RFC3833] Atkins, D., "Threat Analysis of the Domain Name System  
(DNS)", RFC 3833, August 2004.
- [RFC5088] Le Roux, JL., "OSPF Protocol Extensions for Path  
Computation Element (PCE) Discovery", RFC 5088,  
January 2008.
- [RFC5089] Le Roux, JL., "IS-IS Protocol Extensions for Path  
Computation Element (PCE) Discovery", RFC 5089,  
January 2008.
- [RFC5382] Guha, S., "NAT Behavioral Requirements for TCP", RFC 5382,  
October 2008.
- [RFC5452] Hubert, A., "Measures for Making DNS More Resilient  
against Forged Answers", RFC 5452, January 2009.

Authors' Addresses

Qin Wu  
Huawei  
101 Software Avenue, Yuhua District  
Nanjing, Jiangsu 210012  
China

Email: [sunseawq@huawei.com](mailto:sunseawq@huawei.com)

Dhruv Dhody  
Huawei  
Leela Palace  
Bangalore, Karnataka 560008  
INDIA

Email: [dhruv.ietf@gmail.com](mailto:dhruv.ietf@gmail.com)



Network Working Group  
Internet-Draft  
Intended status: Experimental  
Expires: January 14, 2014

F. Zhang, Ed.  
Q. Zhao  
Huawei  
O. Gonzalez de Dios, Ed.  
Telefonica I+D  
R. Casellas  
CTTC  
D. King  
Old Dog Consulting  
July 14, 2013

Extensions to Path Computation Element Communication Protocol (PCEP) for  
Hierarchical Path Computation Elements (PCE)  
draft-zhang-pce-hierarchy-extensions-04

## Abstract

The Hierarchical Path Computation Element (H-PCE) architecture, defined in the companion framework document [RFC6805], provides a mechanism to allow the optimum sequence of domains to be selected, and the optimum end-to-end path to be derived through the use of a hierarchical relationship between domains.

This document defines the Path Computation Element Protocol (PCEP) extensions for the purpose of implementing Hierarchical PCE procedures which are described in the aforementioned document. These extensions are experimental and published for examination, discussion, implementation, and evaluation.

## Status of this Memo

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

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

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

This Internet-Draft will expire on January 14, 2014.

## Copyright Notice

Copyright (c) 2013 IETF Trust and the persons identified as the

document authors. All rights reserved.

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

## Table of Contents

1. Introduction . . . . .	3
1.1. Scope . . . . .	3
1.2. Terminology . . . . .	4
1.3. Requirements Language . . . . .	4
2. Requirements for H-PCE . . . . .	4
2.1. PCEP Requests . . . . .	4
2.1.1. Qualification of PCEP Requests . . . . .	4
2.1.2. Multi-domain Objective Functions . . . . .	5
2.1.3. Multi-domain Metrics . . . . .	6
2.2. Parent PCE Capability Discovery . . . . .	6
2.3. PCE Domain and PCE ID Discovery . . . . .	6
3. PCEP Extensions (Encoding) . . . . .	6
3.1. OPEN Object . . . . .	6
3.1.1. OF Codes . . . . .	6
3.1.2. OPEN Object Flags . . . . .	7
3.1.3. Domain-ID TLV . . . . .	7
3.1.4. PCE-ID TLV . . . . .	9
3.2. RP object . . . . .	9
3.2.1. RP Object Flags . . . . .	9
3.2.2. Domain-ID TLV . . . . .	9
3.3. Metric Object . . . . .	10
3.4. PCEP-ERROR Object . . . . .	10
3.4.1. Hierarchy PCE Error-Type . . . . .	10
3.5. NO-PATH Object . . . . .	10
4. H-PCE Procedures . . . . .	10
4.1. OPEN Procedure between Child PCE and Parent PCE . . . . .	11
4.2. Procedure to Obtain Domain Sequence . . . . .	11
5. Error Handling . . . . .	11
6. Manageability Considerations . . . . .	12
7. IANA Considerations . . . . .	12
8. Security Considerations . . . . .	12
9. Contributing Authors . . . . .	12
10. Acknowledgments . . . . .	12
11. Normative References . . . . .	13
Authors' Addresses . . . . .	13

## 1. Introduction

[RFC6805] describes a Hierarchical PCE (H-PCE) architecture which can be used for computing end-to-end paths for inter-domain MPLS Traffic Engineering (TE) and GMPLS Label Switched Paths (LSPs).

Within the hierarchical PCE architecture, the parent PCE is used to compute a multi-domain path based on the domain connectivity information. A child PCE may be responsible for a single domain or multiple domains, it is used to compute the intra-domain path based on its domain topology information.

The H-PCE end-to-end domain path computation procedure is described below:

- o A path computation client (PCC) sends the inter-domain path computation requests to the child PCE responsible for its domain;
- o The child PCE forwards the request to the parent PCE;
- o The parent PCE computes the likely domain paths from the ingress domain to the egress domain;
- o The parent PCE sends the intra-domain path computation requests (between the domain border nodes) to the child PCEs which are responsible for the domains along the domain path;
- o The child PCEs return the intra-domain paths to the parent PCE;
- o The parent PCE constructs the end-to-end inter-domain path based on the intra-domain paths;
- o The parent PCE returns the inter-domain path to the child PCE;
- o The child PCE forwards the inter-domain path to the PCC.

In addition, the parent PCE may be requested to provide only the sequence of domains to a child PCE so that alternative inter-domain path computation procedures, including Per Domain (PD) [RFC5152] and Backwards Recursive Path Computation (BRPC) [RFC5441] may be used.

This document defines the PCEP extensions for the purpose of implementing Hierarchical PCE procedures, which are described in [RFC6805].

### 1.1. Scope

The following functions are out of scope of this document.

- o Finding end point addresses;
- o Parent Traffic Engineering Database (TED) methods;
- o Domain connectivity;

The document also uses a number of [editor notes] to describe options and alternative solutions. These options and notes will be removed before publication once agreement is reached.

## 1.2. Terminology

This document uses the terminology defined in [RFC4655], [RFC5440] and the additional terms defined in section 1.4 of [RFC6805].

## 1.3. Requirements Language

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

## 2. Requirements for H-PCE

This section compiles the set of requirements of the PCEP protocol to support the H-PCE architecture and procedures.

[RFC6805] identifies high-level requirements of PCEP extensions required to support the hierarchical PCE model.

### 2.1. PCEP Requests

The PCReq messages are used by a PCC or PCE to make a path computation request to a PCE. In order to achieve the full functionality of the H-PCE procedures, the PCReq message needs to include:

- o Qualification of PCE Requests.
- o Multi-domain Objective Functions (OF).
- o Multi-domain Metrics.

#### 2.1.1. Qualification of PCEP Requests

As described in section 4.8.1 of [RFC6805], the H-PCE architecture introduces new request qualifications, which are:



- o It MUST be possible for a child PCE to indicate that a request it sends to a parent PCE should be satisfied by a domain sequence only, that is, not by a full end-to-end path. This allows the child PCE to initiate a per-domain (PD) [RFC5152] or a backward recursive path computation (BRPC) [RFC5441].
- o As stated in [RFC6805], section 4.5, if a PCC knows the egress domain, it can supply this information as the path computation request. It SHOULD be possible to specify the destination domain information in a PCEP request, if it is known.

#### 2.1.2. Multi-domain Objective Functions

For inter-domain path computation, there are two new objective functions which are defined in section 1.3.1 and 4.1 of [RFC6805]:

- o Minimize the number of domains crossed. A domain can be either an Autonomous System (AS) or an Internal Gateway Protocol (IGP) area depending on the type of multi-domain network hierarchical PCE is applied to.
- o Disallow domain re-entry.[Editor's note: Disallow domain re-entry may not be an objective function, but an option in the request].

During the PCEP session establishment procedure, the parent PCE needs to be capable of indicating the Objective Functions (OF) capability in the Open message. This capability information may then be announced by child PCEs, and used for selecting the PCE when a PCC wants a path that satisfies one or multiple inter-domain objective functions.

When a PCC requests a PCE to compute an inter-domain path, the PCC needs also to be capable of indicating the new objective functions for inter-domain path. Note that a given child PCE may also act as a parent PCE.

For the reasons described previously, new OF codes need to be defined for the new inter-domain objective functions. Then the PCE can notify its new inter-domain objective functions to the PCC by carrying them in the OF-list TLV which is carried in the OPEN object. The PCC can specify which objective function code to use, which is carried in the OF object when requesting a PCE to compute an inter-domain path.

The proposed solution may need to differentiate between the OF code that is requested at the parent level, and the OF code that is requested at the intra-domain (child domain).

A parent PCE MUST be capable of ensuring homogeneity, across domains, when applying OF codes for strict OF intra-domain requests.

#### 2.1.3. Multi-domain Metrics

For inter-domain path computation, there are several path metrics of interest [Editor's note: Current framework only mentions metric objectives. The metric itself should be also defined]:

- o Domain count (number of domains crossed).
- o Border Node count.

A PCC may be able to limit the number of domains crossed by applying a limit on these metrics.

#### 2.2. Parent PCE Capability Discovery

Parent and child PCE relationships are likely to be configured. However, as mentioned in [RFC6805], it would assist network operators if the child and parent PCE could indicate their H-PCE capabilities.

During the PCEP session establishment procedure, the child PCE needs to be capable of indicating to the parent PCE whether it requests the parent PCE capability or not. Also, during the PCEP session establishment procedure, the parent PCE needs to be capable of indicating whether its parent capability can be provided or not.

#### 2.3. PCE Domain and PCE ID Discovery

A PCE domain is a single domain with an associated PCE. Although it is possible for a PCE to manage multiple domains. The PCE domain may be an IGP area or AS.

The PCE ID is an IPv4 and/or IPv6 address that is used to reach the parent/child PCE. It is RECOMMENDED to use an address that is always reachable if there is any connectivity to the PCE.

The PCE ID information and PCE domain identifiers may be provided during the PCEP session establishment procedure or the domain connectivity information collection procedure.

### 3. PCEP Extensions (Encoding)

#### 3.1. OPEN object

##### 3.1.1. OF Codes

This H-PCE experiment will be carried out using the following OF codes:

- o MTD
  - \* Name: Minimize the number of Transit Domains.
  - \* Objective Function Code.
  - \* Description: Find a path P such that it passes through the lnumber of transit domains.
- o MBN
  - \* Name: Minimize the number of border nodes.
  - \* Objective Function Code.
  - \* Description: Find a path P such that it passes through the least number of border nodes.
- o DDR
  - \* Name: Disallow Domain Re-entry (DDR)
  - \* Objective Function Code.
  - \* Description: Find a path P such that does not entry a domain more than once.

### 3.1.2. OPEN Object Flags

This H-PCE experiment will also require two OPEN object flags:

- o Parent PCE Request bit (to be assigned by IANA, recommended bit 0): if set, it would signal that the child PCE wishes to use the peer PCE as a parent PCE.
- o Parent PCE Indication bit (to be assigned by IANA, recommended bit 1): if set, it would signal that the PCE can be used as a parent PCE by the peer PCE.

### 3.1.3. Domain-ID TLV

The Domain-ID TLV for this H-PCE experiment is defined below:

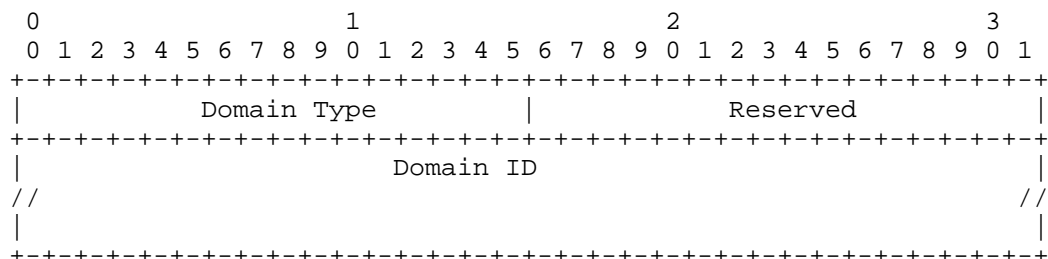


Figure 1: Domain-ID TLV

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

- o Type=1: the Domain ID field carries an IGP Area ID.
- o Type=2: the Domain ID field carries an AS number.

Domain ID (variable): Indicates an IGP Area ID or AS number. It can be 2 bytes, 4 bytes or 8 bytes long depending on the domain identifier used.

[Editor's note: draft-dhody-pce-pcep-domain-sequence, section 3.2 deals with the encoding of domain sequences, using ERO-subobjects. Work is ongoing to define domain identifiers for OSPF-TE areas, IS-IS area (which are variable sized), 2-byte and 4-byte AS number, and any other domain that may be defined in the future. It uses RSVP-TE subobject discriminators, rather than new type 1/ type 2. A domain sequence may be encoded as a route object. The "VALUE" part of the TLV could follow common RSVP-TE subobject format:

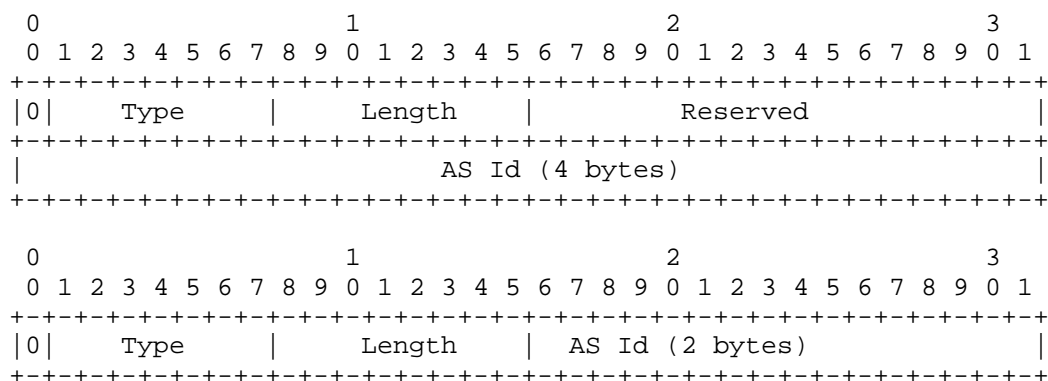


Figure 2: Alternative Domain-ID TLV

### 3.1.4. PCE-ID TLV

The type of PCE-ID TLV for this H-PCE experiment is defined below:

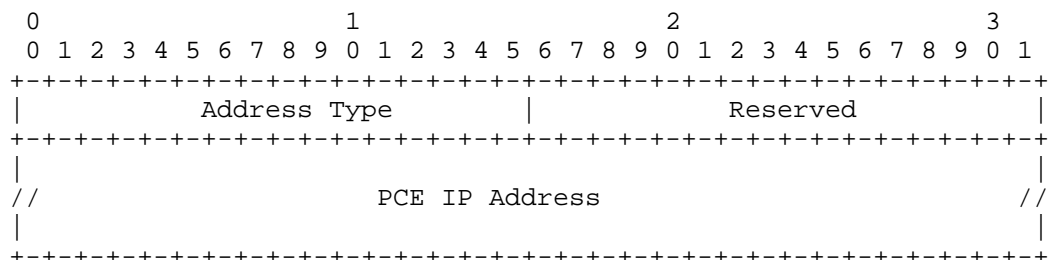


Figure 3: PCE-ID TLV

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

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

[Editor's note: [RFC5886] already defines the PCE-ID object. If a semantically equivalent PCE-ID TLV is needed (to avoid modifying message grammars to include the object), it can align with the PCEP object: in any case, the length (4 / 16 bytes) can be used to know whether it is an IPv4 or an IPv6 PCE, the address type is not needed.]

### 3.2. RP object

### 3.2.1. RP Object Flags

The following RP object flags are defined for this H-PCE experiment:

- o Domain Path Request bit: if set, it means the child PCE wishes to get the domain sequence.
- o Destination Domain Query bit: if set, it means the parent PCE wishes to get the destination domain ID.

### 3.2.2. Domain-ID TLV

The format of this TLV is defined in Section 3.1.3. This TLV can be carried in an OPEN object to indicate a (list of) managed domains, or carried in a RP object to indicate the destination domain ID when a child PCE responds to the parent PCE's destination domain query by a PCRep message.

[Editors note. In some cases, the Parent PCE may need to allocate a node which is not necessarily the destination node.]

### 3.3. Metric Object

There are two new metrics defined in this document for H-PCE:

- o Domain count (number of domains crossed).
- o Border Node Count (number of border nodes crossed).

### 3.4. PCEP-ERROR object

#### 3.4.1. Hierarchy PCE Error-Type

A new PCEP Error-Type is used for this H-PCE experiment and is defined below:

Error-Type	Meaning
19	H-PCE error Error-value=1: parent PCE capability cannot be provided

H-PCE error table

### 3.5. NO-PATH Object

To communicate the reason(s) for not being able to find a multi-domain path or domain sequence, the NO-PATH object can be used in the PCRep message. [RFC5440] defines the format of the NO-PATH object. The object may contain a NO-PATH-VECTOR TLV to provide additional information about why a (domain) path computation has failed.

Three new bit flags are defined to be carried in the Flags field in the NO-PATH-VECTOR TLV carried in the NO-PATH Object.

- o Bit 23: When set, the parent PCE indicates that destination domain unknown;
- o Bit 22: When set, the parent PCE indicates unresponsive child PCE(s);
- o Bit 21: When set, the parent PCE indicates no available resource available in one or more domain(s).

## 4. H-PCE Procedures

#### 4.1. OPEN Procedure between Child PCE and Parent PCE

If a child PCE wants to use the peer PCE as a parent, it can set the parent PCE request bit in the OPEN object carried in the Open message during the PCEP session creation procedure. If the peer PCE does not want to provide the parent function to the child PCE, it must send a PCErr message to the child PCE and clear the parent PCE indication bit in the OPEN object.

If the parent PCE can provide the parent function to the peer PCE, it may set the parent PCE indication bit in the OPEN object carried in the Open message during the PCEP session creation procedure.

The PCE may also report its PCE ID and list of domain ID to the peer PCE by specifying them in the PCE-ID TLV and List of Domain-ID TLVs in the OPEN object carried in the Open message during the PCEP session creation procedure.

The OF codes defined in this document can be carried in the OF-list TLV of the OPEN object. If the OF-list TLV carries the OF codes, it means that the PCE is capable of implementing the corresponding objective functions. This information can be used for selecting a proper parent PCE when a child PCE wants to get a path that satisfies a certain objective function.

When a specific child PCE sends a PCReq to a peer PCE that requires parental activity and the peer PCE does not want to act as the parent for it, the peer PCE should send a PCErr message to the child PCE and specify the error-type (IANA) and error-value (1) in the PCEP-ERROR object.

#### 4.2. Procedure to obtain Domain Sequence

If a child PCE only wants to get the domain sequence for a multi-domain path computation from a parent PCE, it can set the Domain Path Request bit in the RP object carried in a PCReq message. The parent PCE which receives the PCReq message tries to compute a domain sequence for it. If the domain path computation succeeds the parent PCE sends a PCRep message which carries the domain sequence in the ERO to the child PCE. The domain sequence is specified as AS or AREA ERO sub-objects (type 32 for AS [RFC3209] or a to-be-defined IGP area type). Otherwise it sends a PCReq message which carries the NO-PATH object to the child PCE.

### 5. Error Handling

A PCE that is capable of acting as a parent PCE might not be configured or willing to act as the parent for a specific child PCE.

This fact could be determined when the child sends a PCReq that requires parental activity (such as querying other child PCEs), and could result in a negative response in a PCEP Error (PCErr) message and indicate the hierarchy PCE error types.

Additionally, the parent PCE may fail to find the multi-domain path or domain sequence due to one or more of the following reasons:

- o A child PCE cannot find a suitable path to the egress;
- o The parent PCE do not hear from a child PCE for a specified time;
- o The objective functions specified in the path request cannot be met.

In this case, the parent PCE MAY need to send a negative path computation reply specifying the reason. This can be achieved by including NO-PATH object in the PCRep message. Extension to NO-PATH object is needed to include the aforementioned reasons.

## 6. Manageability Considerations

TBD.

## 7. IANA Considerations

Due to the experimental nature of this draft no IANA requests are made.

## 8. Security Considerations

To be added.

## 9. Contributing Authors

Xian Zhang  
Huawei  
zhang.xian@huawei.com

## 10. Acknowledgments

To be added.



## 11. Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC3209] Awduche, D., Berger, L., Gan, D., Li, T., Srinivasan, V., and G. Swallow, "RSVP-TE: Extensions to RSVP for LSP Tunnels", RFC 3209, December 2001.
- [RFC4655] Farrel, A., Vasseur, J., and J. Ash, "A Path Computation Element (PCE)-Based Architecture", RFC 4655, August 2006.
- [RFC5152] Vasseur, JP., Ayyangar, A., and R. Zhang, "A Per-Domain Path Computation Method for Establishing Inter-Domain Traffic Engineering (TE) Label Switched Paths (LSPs)", RFC 5152, February 2008.
- [RFC5440] Vasseur, JP. and JL. Le Roux, "Path Computation Element (PCE) Communication Protocol (PCEP)", RFC 5440, March 2009.
- [RFC5441] Vasseur, JP., Zhang, R., Bitar, N., and JL. Le Roux, "A Backward-Recursive PCE-Based Computation (BRPC) Procedure to Compute Shortest Constrained Inter-Domain Traffic Engineering Label Switched Paths", RFC 5441, April 2009.
- [RFC5886] Vasseur, JP., Le Roux, JL., and Y. Ikejiri, "A Set of Monitoring Tools for Path Computation Element (PCE)-Based Architecture", RFC 5886, June 2010.
- [RFC6805] King, D. and A. Farrel, "The Application of the Path Computation Element Architecture to the Determination of a Sequence of Domains in MPLS and GMPLS", RFC 6805, November 2012.

## Authors' Addresses

Fatai Zhang (editor)  
Huawei  
Huawei Base, Bantian, Longgang District  
Shenzhen, 518129  
China

Phone: +86-755-28972912  
Email: zhangfatai@huawei.com

Quintin Zhao  
Huawei  
125 Nagog Technology Park  
Acton, MA 01719  
US

Phone:  
Email: qzhao@huawei.com

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

Phone: +34913128832  
Email: ogondio@tid.es

Ramon Casellas  
CTTC  
Av. Carl Friedrich Gauss n.7  
Castelldefels, Barcelona  
Spain

Phone: +34 93 645 29 00  
Email: ramon.casellas@cttc.es

Daniel King  
Old Dog Consulting  
UK

Phone:  
Email: daniel@olddog.co.uk



Network Working Group  
Internet-Draft  
Intended status: Informational  
Expires: November 25, 2013

X. Zhang, Ed.  
Huawei Technologies  
I. Minei, Ed.  
Juniper Networks, Inc.  
May 24, 2013

Applicability of Stateful Path Computation Element (PCE)  
draft-zhang-pce-stateful-pce-app-04

Abstract

A stateful Path Computation Element (PCE) maintains information about Label Switched Path (LSP) characteristics and resource usage within a network in order to provide traffic engineering calculations for its associated Path Computation Clients (PCCs). This document describes general considerations for a stateful PCE deployment and examines its applicability and benefits through a number of use cases. Path Computation Element Protocol (PCEP) extensions required for stateful PCE usage are covered in separate documents.

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 November 25, 2013.

Copyright Notice

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

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

to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

## Table of Contents

1. Introduction . . . . .	3
2. Terminology . . . . .	3
3. Overview of stateful PCE . . . . .	4
4. Deployment considerations . . . . .	5
4.1. Multi-PCE deployments . . . . .	5
4.2. LSP State Synchronization . . . . .	5
4.3. PCE Survivability . . . . .	5
5. Application scenarios . . . . .	6
5.1. Optimization of LSP placement . . . . .	6
5.1.1. Throughput Maximization and Bin Packing . . . . .	7
5.1.2. Deadlock . . . . .	8
5.1.3. Minimum Perturbation . . . . .	10
5.1.4. Predictability . . . . .	11
5.2. Auto-bandwidth Adjustment . . . . .	12
5.3. Bandwidth Scheduling . . . . .	13
5.4. Recovery . . . . .	13
5.4.1. Protection . . . . .	13
5.4.2. Restoration . . . . .	15
5.4.3. SRLG Diversity . . . . .	16
5.5. Maintenance of Virtual Network Topology (VNT) . . . . .	16
5.6. LSP Re-optimization . . . . .	17
5.7. Resource Defragmentation . . . . .	17
5.8. Impairment-Aware Routing and Wavelength Assignment (IA-RWA) . . . . .	18
6. Security Considerations . . . . .	19
7. Contributing Authors . . . . .	19
8. Acknowledgements . . . . .	21
9. References . . . . .	21
9.1. Normative References . . . . .	21
9.2. Informative References . . . . .	21
Appendix A. Editorial notes and open issues . . . . .	23
Authors' Addresses . . . . .	23

## 1. Introduction

[RFC4655] defines the architecture for a Path Computation Element (PCE)-based model for the computation of Multiprotocol Label Switching (MPLS) and Generalized MPLS (GMPLS) Traffic Engineering Label Switched Paths (TE LSPs). To perform such a constrained computation, a PCE stores the network topology (i.e., TE links and nodes) and resource information (i.e., TE attributes) in its TE Database (TED). [RFC5440] describes the Path Computation Element Protocol (PCEP). PCEP defines the communication between a Path Computation Client (PCC) and a Path Control Element (PCE), or between two PCEs, enabling computation of Multiprotocol Label Switching (MPLS) for Traffic Engineering Label Switched Path (TE LSP) characteristics. Extensions for support of GMPLS in PCEP are defined in [I-D.ietf-pce-gmpls-pcep-extensions].

As per [RFC4655], a PCE can be either stateful or stateless. Stateless PCEs have been shown to be useful in many scenarios, including constraint-based path computation in multi-domain/multi-layer networks. Compared to a stateless PCE, a stateful PCE has access to not only the network state, but also to the set of active paths and their reserved resources. Furthermore, a stateful PCE might also retain information regarding LSPs under construction in order to reduce churn and resource contention. This state allows the PCE to compute constrained paths while considering individual LSPs and their interactions. Note that this requires reliable state synchronization mechanisms between the PCE and the network, PCE and PCC, and between cooperating PCEs, with potentially significant control plane overhead and maintenance of a large amount of state data, as explained in [RFC4655].

This document describes how a stateful PCE can be used to solve various problems for MPLS-TE and GMPLS networks, and the benefits it brings to such deployments. Note that alternative solutions relying on stateless PCEs may also be possible for some of these use cases, and will be mentioned for completeness where appropriate.

## 2. Terminology

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

This document uses the following terms defined in [I-D.ietf-pce-stateful-pce]: Passive Stateful PCE, Active Stateful PCE, Delegation, Revocation, Delegation Timeout Interval, LSP State Report, LSP Update Request, LSP State Database.

This document defines the following term:

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

### 3. Overview of stateful PCE

This section is included for the convenience of the reader, please refer to the referenced documents for details of the operation.

[I-D.ietf-pce-stateful-pce] specifies a set of extensions to PCEP to enable stateful control of tunnels within and across PCEP sessions in compliance with [RFC4657]. It includes mechanisms to effect tunnel state synchronization between PCCs and PCEs, delegation of control over tunnels to PCEs, and PCE control of timing and sequence of path computations within and across PCEP sessions.

[I-D.ietf-pce-stateful-pce] applies equally to MPLS-TE and GMPLS LSPs.

Several new functions were added in PCEP to support stateful PCEs and are described in [I-D.ietf-pce-stateful-pce]. A function can be initiated either from a PCC towards a PCE (C-E) or from a PCE towards a PCC (E-C). The new functions are:

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

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

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

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

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

[I-D.sivabalan-pce-disco-stateful] defines the extensions needed to support autodiscovery of stateful PCEs when using the IGPs for PCE discovery.

#### 4. Deployment considerations

This section discusses generic issues with Stateful PCE deployments, and how specific protocol mechanisms can be used to address them.

##### 4.1. Multi-PCE deployments

Stateless and stateful PCEs can co-exist in the same network and be in charge of path computation of different types. To solve the problem of distinguishing between the two types of PCEs, either discovery or configuration may be used. The capability negotiation in [I-D.ietf-pce-stateful-pce] ensures correct operation when the PCE address is configured on the PCC.

##### 4.2. LSP State Synchronization

A stateful PCE maintains two sets of information for use in path computation. The first is the Traffic Engineering Database (TED) which includes the topology and resource state in the network. This information can be obtained by a stateful PCE using the same mechanisms as a stateless PCE (see [RFC4655]). The second is the LSP State Database (LSP-DB), in which a PCE stores attributes of all active LSPs in the network, such as their paths through the network, bandwidth/resource usage, switching types and LSP constraints. The stateful PCE extensions defined in [I-D.ietf-pce-stateful-pce] support population of this database using information received from the network nodes via LSP State Report messages. Population of the LSP database via other means is not precluded.

##### 4.3. PCE Survivability

For a stateful PCE, an important issue is to get the LSP state information resynchronized after a restart. [I-D.ietf-pce-stateful-pce] includes support of a synchronization function, allowing the PCC to synchronize its LSP state with the PCE. This can be applied equally to an Label Edge Router (LER) client or another PCE, allowing for support of multiple ways of re-acquiring



the LSP database on a restart. For example, the state can be retrieved from the network nodes, or from another stateful PCE. Because synchronization may also be skipped, if a PCE implementation has the means to retrieve its database in a different way (for example from a backup copy stored locally), the state can be restored without further overhead in the network. Note that locally recovering the state would still require some degree of resynchronization to ensure that the recovered state is indeed up-to-date.

## 5. Application scenarios

In the following sections, several use cases are described, showcasing scenarios that benefit from the deployment of a stateful PCE.

### 5.1. Optimization of LSP placement

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

Some of the use cases below are focused on MPLS-TE deployments, but may also apply to GMPLS. Unless otherwise cited, use cases assume that all LSPs listed exist at the same LSP priority.

The main benefit in the cases below comes from moving away from an asynchronous PCC-driven mode of operation to a model that allows for central control over LSP computations and setup, and focuses specifically on the active stateful PCE model of operation.

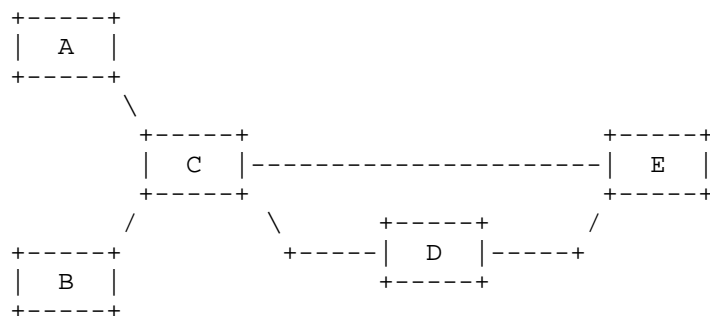


Figure 1: Reference topology 1

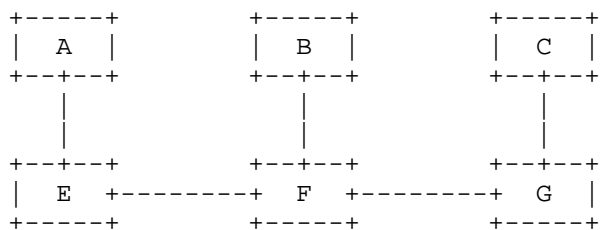


Figure 2: Reference topology 2

#### 5.1.1. Throughput Maximization and Bin Packing

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

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

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

Table 1: Link parameters for Throughput use case

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

Table 2: Throughput use case demand time series

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

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

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

Table 3: Link parameters for Bin Packing use case

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

Table 4: Bin Packing use case demand time series

#### 5.1.2. Deadlock

This section discusses a use case of cross-LSP impact under degraded operation. Most existing RSVP-TE implementations will not tear down established LSPs in the event of the failure of the bandwidth

increase procedure detailed in [RFC3209]. This behavior is directly implied to be correct in [RFC3209] and is often desirable from an operator's perspective, because either a) the destination prefixes are not reachable via any means other than MPLS or b) this would result in significant packet loss as demand is shifted to other LSPs in the overlay mesh.

In addition, there are currently few implementations offering dynamic ingress admission control (policing of the traffic volume mapped onto an LSP) at the LER. Having ingress admission control on a per LSP basis is not necessarily desirable from an operational perspective, as a) one must over-provision tunnels significantly in order to avoid deleterious effects resulting from stacked transport and flow control systems and b) there is currently no efficient commonly available northbound interface for dynamic configuration of per LSP ingress admission control (such an interface could easily be defined using the extensions for stateful PCE, but has not been yet at the time of this writing).

Lack of ingress admission control coupled with the behavior in [RFC3209] may result in LSPs operating out of profile for significant periods of time. It is reasonable to expect that these out-of-profile LSPs will be operating in a degraded state and experience traffic loss, but because they end up sharing common network interfaces with other LSPs operating within their bandwidth reservations, they will end up impacting the operation of the in-profile LSPs, even when there is unused network capacity elsewhere in the network. Furthermore, this behavior will cause information loss in the TED with regards to the actual available bandwidth on the links used by the out-of-profile LSPs, as the reservations on the links no longer reflect the capacity used.

Reference Topology 1 in Figure 1 and Tables 5 and 6 show a use case that demonstrates this behavior. Two LSPs, LSP 1 and LSP 2 are signaled with demand 2 and routed along paths A-C-D-E and B-C-D-E respectively. At a later time, the demand of LSP 1 increases to 20. Under such a demand, the LSP cannot be resigaled. However, the existing LSP will not be torn down. In the absence of ingress policing, traffic on LSP 1 will cause degradation for traffic of LSP 2 (due to oversubscription on the links C-D and D-E), as well as information loss in the TED with regard to the actual network state.

The problem could be easily ameliorated by global visibility of LSP state coupled with PCC-external demand measurements and placement of two LSPs on disjoint links. Note that while the demand of 20 for LSP 1 could never be satisfied in the given topology, what could be achieved would be isolation from the ill-effects of the (unsatisfiable) increased demand.

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

Table 5: Link parameters for the 'Degraded operation' example

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

Table 6: Degraded operation demand time series

#### 5.1.3. Minimum Perturbation

As a result of both the lack of visibility into global LSP state and the lack of control over event ordering across PCE sessions, unnecessary perturbations may be introduced into the network by a stateless PCE. Tables 7 and 8 show an example of an unnecessary network perturbation using Reference Topology 1 in Figure 1. In this case an unimportant (high LSP priority value) LSP (LSP1) is first set up along the shortest path. At time 2, which is assumed to be relatively close to time 1, a second more important (lower LSP-priority value) LSP (LSP2) is established, preempting LSP1, potentially causing traffic loss. LSP1 is then reestablished on the longer A-C-E path.

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

Table 7: Link parameters for the 'Minimum-Perturbation' example

Time	LSP	Src	Dst	Demand	LSP Prio	Routable	Path
1	1	A	E	7	7	Yes	A-C-D-E
2	2	B	E	7	0	Yes	B-C-D-E
3	1	A	E	7	7	Yes	A-C-E

Table 8: Minimum-Perturbation LSP and demand time series

A stateful PCE can help in this scenario by evaluating both requests at the same time (due to their proximity in time). This will ensure placement of the more important LSP along the shortest path, avoiding the preemption of the lower priority LSP.

#### 5.1.4. Predictability

Randomization of reservation events caused by lack of control over event ordering across PCE sessions results in poor predictability in LSP routing. An offline system applying a consistent optimization method will produce predictable results to within either the boundary of forecast error when reservations are over-provisioned by reasonable margins or to the variability of the signal and the forecast error when applying some hysteresis in order to minimize churn. Predictable results are valuable for being able to simulate the network and reliably test it under various scenarios, especially under various failure modes and planned maintenances when predictable path characteristics are desired under contention for network resources.

Reference Topology 1 and Tables 9, 10 and 11 show the impact of event ordering and predictability of LSP routing.

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

Table 9: Link parameters for the 'Predictability' example

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

Table 10: Predictability LSP and demand time series 1

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

Table 11: Predictability LSP and demand time series 2

As can be shown in the example, both LSPs were routed in both cases, but along very different paths. This would be a challenge if reliable simulation of the network was attempted. A stateful PCE can solve this through control over LSP ordering.

## 5.2. Auto-bandwidth Adjustment

The bandwidth requirement of LSPs often change over time, requiring resizing the LSP. Currently the head-end node performs this function by monitoring the actual bandwidth usage, triggering a recomputation and resignaling when a threshold is reached. This operation is referred as auto-bandwidth adjustment. The head-end node either recomputes the path locally, or it requests a recomputation from a PCE by sending a PCReq message. In the latter case, the PCE computes a new path and provides the new route suggestion. Upon receiving the reply from the PCE, the PCC re-signals the LSP in Shared-Explicit (SE) mode along the newly computed path. If a passive stateful PCE is used, only the new bandwidth information is needed to trigger a path re-computation since the LSP information is already known to the PCE. Note that in this scenario, the head-end node is the one that drives the LSP resizing based on local information, and that the difference between using a stateless and a passive stateful PCE is in the level of optimization of the LSP placement as discussed in the previous section.

A more interesting smart bandwidth adjustment case is one where the LSP resizing decision is done by an external entity, with access to additional information such as historical trending data, application-specific information about expected demands or policy information, as well as knowledge of the actual desired flow volumes. In this case

an active stateful PCE provides an advantage in both the computation with knowledge of all LSPs in the domain and in the ability to trigger bandwidth modification of the LSP.

### 5.3. Bandwidth Scheduling

Bandwidth scheduling allows network operators to reserve resources in advance according to the agreements with their customers, and allow them to transmit data with specified starting time and duration, for example for a scheduled bulk data replication between data centers.

Traditionally, this can be supported by NMS operation through path pre-establishment and activation on the agreed starting time. 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. It can also be accomplished through GMPLS protocol extensions by carrying the related request information (e.g., starting time and duration) across the network. Nevertheless, this method inevitably increases the complexity of signaling and routing process.

A passive stateful PCE can support this application with better efficiency since it can alleviate the burden of processing on network elements. This requires the PCE to maintain the scheduled LSPs and their associated resource usage, as well as the ability of head-ends to trigger signaling for LSP setup/deletion at the correct time. This approach requires coarse time synchronization between PCEs and PCCs. If an active stateful PCE is available, the PCE can trigger the setup/deletion of scheduled requests in a centralized manner, without modification of existing head-end behaviors.

### 5.4. Recovery

The recovery use cases discussed in the following sections show how leveraging a stateful PCE can simplify the computation of recovery path(s). In particular, two characteristics of a stateful PCE are used: 1) using information stored in the LSP-DB for determining shared protection resources and 2) performing computations with knowledge of all LSPs in a domain.

#### 5.4.1. Protection

For protection purposes, a PCC may send a request to a PCE for computing a set of paths for a given LSP. Alternatively, the PCC can send multiple requests to the PCE, asking for working and backup LSPs separately. Either way, the resources bound to backup paths can be shared by different LSPs to improve the overall network efficiency, such as m:n protection or pre-configured shared mesh recovery



techniques as specified in [RFC4427]. If resource sharing is supported for LSP protection, the information relating to existing LSPs is required to avoid allocation of shared protection resources to two LSPs that might fail together and cause protection contention issues. A stateless PCE can accommodate this use case by having the PCC pass in this information as a constraint to the path computation request. A stateful PCE can more easily accommodate this need using the information stored in its LSP-DB.

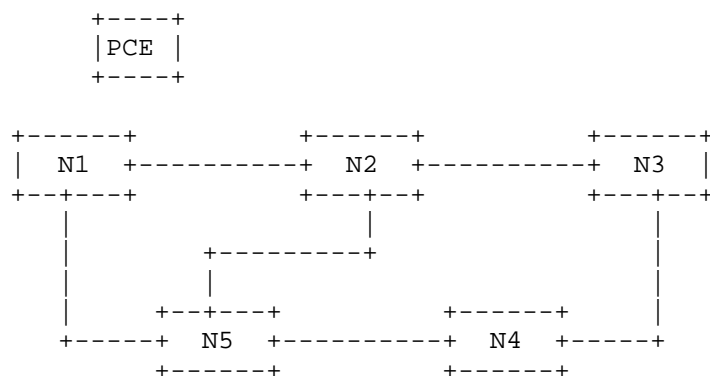


Figure 3: Reference topology 3

For example, in the network depicted in Figure 3, suppose there exists LSP1 with working path LSP1\_working following N1->N5 and with backup path LSP1\_backup following N1->N2->N5. A request arrives asking for a working and backup path pair to be computed for LSP2, for a request from N2 to N5. If the PCE decides LSP2\_working follows N2->N1->N5, then the backup path LSP2\_backup should not use the same protection resource with LSP1 since LSP2 shares part of its resource (specifically N1->N5) with LSP1 (i.e., these two LSPs are in the same shared risk group). Alternatively, there is no such constraint if N2->N3->N4->N5 is chosen for LSP2\_working.

If a stateless PCE is used, the head node N2 needs to be aware of the existence of LSPs which share the route of LSP2\_working and of the details of their protection resources. N2 must pass this information to the PCE as a constraint so as to request a path with SRLG diversity. On the other hand, a stateful PCE can get the LSPs information by itself and can achieve the goal of finding SRLG-diversified protection paths for both LSPs. This is made possible by comparing the LSP resource usage exploiting the LSP DB accessible by the stateful PCE.

#### 5.4.2. Restoration

In case of a link failure, such as fiber cut, multiple LSPs may fail at the same time. Thus, the source nodes of the affected LSPs will be informed of the failure by the nodes detecting the failure. These source nodes will send requests to a PCE for rerouting. In order to reuse the resource taken by an existing LSP, the source node can send a PCReq message including the XRO object with F bit set, together with RRO object, as specified in [RFC5521].

If a stateless PCE is exploited, it might respond to the rerouting requests separately if they arrive at different times. Thus, it might result in sub-optimal resource usage. Even worse, it might unnecessarily block some of the rerouting requests due to insufficient resources for later-arrived rerouting messages. If a stateful PCE is used to fulfill this task, it can re-compute the affected LSPs concurrently while reusing part of the existing LSPs resources when it is informed of the failed link identifier provided by the first request. This is made possible since the stateful PCE can check what other LSPs are affected by the failed link and their route information by inspecting its LSP-DB. As a result, a better performance, such as better resource usage, minimal probability of blocking upcoming new rerouting requests sent as a result of the link failure, can be achieved.

In order to further reduce the amount of LSP rerouting messages flow in the network, the notification can be performed at the node(s) which detect the link failure. For example, suppose there are two LSPs in the network as shown in Figure 3: (i) LSP1: N1->N5->N4->N3; (ii) LSP2: N2->N5->N4. They traverse the failed link between N5-N4. When N4 detects the failure, it can send a notification message to a stateful PCE. Note that the stateful PCE stores the path information of the LSPs that are affected by the link failure, so it does not need to acquire this information from N4. Moreover, it can make use of the bandwidth resources occupied by the affected LSPs when performing path recalculation. After N4 receives the new paths from the PCE, it notifies the ingress nodes of the LSPs, i.e., N1 and N2, and specifies the new paths which should be used as the rerouting paths. To support this, it would require extensions to the existing signaling protocols.

Alternatively, if the target is to avoid resource contention within the time-window of high LSP requests, a stateful PCE can retain the under-construction LSP resource usage information for a given time and exclude it from being used for forthcoming LSPs request. In this way, it can ensure that the resource will not be double-booked and thus the issue of resource contention and computation crank-backs can be resolved.

#### 5.4.3. SRLG Diversity

An alternative way to achieve efficient resilience is to maintain SRLG disjointness between LSPs, irrespective of whether these LSPs share the source and destination nodes or not. This can be achieved at provisioning time, if the routes of all the LSPs are requested together, using a synchronized computation of the different LSPs with SRLG disjointness constraint. If the LSPs need to be provisioned at different times (more general, the routes are requested at different times, e.g. in the case of a restoration), the PCC can specify, as constraints to the path computation a set of Shared Risk Link Groups (SRLGs) using the Explicit Route Object [RFC5521]. However, for the latter to be effective, it is needed that the entity that requests the route to the PCE maintains updated SRLG information of all the LSPs to which it must maintain the disjointness. A stateless PCE can compute an SRLG-disjoint path by inspecting the TED and precluding the links with the same SRLG values specified in the PCReq message sent by a PCC.

A stateful PCE maintains the updated SRLG information of the established LSPs in a centralized manner. Therefore, the PCC can specify as constraints to the path computation the SRLG disjointness of a set of already established LSPs by only providing the LSP identifiers.

#### 5.5. Maintenance of Virtual Network Topology (VNT)

In Multi-Layer Networks (MLN), a Virtual Network Topology (VNT) [RFC5212] consists of a set of one or more TE LSPs in the lower layer which provides TE links to the upper layer. In [RFC5623], the PCE-based architecture is proposed to support path computation in MLN networks in order to achieve inter-layer TE.

The establishment/teardown of a TE link in VNT needs to take into consideration the state of existing LSPs and/or new LSP request(s) in the higher layer. As specified in [RFC5623], a VNT manager (VNTM) is in charge of setting up connections in the lower layer to provide TE links for upper layer. Hence, when a stateless PCE cannot find the route for a request based on the upper layer topology information, it needs to interact with the VNTM and rely on the VNTM to decide whether to set up or remove a TE link or not. On the other hand, a stateful PCE can make the decision of when and how to modify the VNT either to accommodate new LSP requests or to re-optimize resource usage across layers irrespective of the PCE models as described in [RFC5623].

## 5.6. LSP Re-optimization

In order to make efficient usage of network resources, it is sometimes desirable to re-optimize one or more LSPs dynamically. In the case of a stateless PCE, in order to optimize network resource usage dynamically through online planning, a PCC must send a request to the PCE together with detailed path/bandwidth information of the LSPs that need to be concurrently optimized. This means the PCC must be able to determine when and which LSPs should be optimized. In the case of a stateful PCE, given the LSP state information in the LSP database, the process of dynamic optimization of network resources can be automated without requiring the PCC to supply LSP state information or to trigger the request. Moreover, since a stateful PCE can maintain information for all LSPs that are in the process of being set up and since it may have the ability to control timing and sequence of LSP setup/deletion, the optimization procedures can be performed more intelligently and effectively.

A special case of LSP re-optimization is Global Concurrent Optimization (GCO) [RFC5557]. Global control of LSP operation sequence in [RFC5557] is predicated on the use of what is effectively a stateful (or semi-stateful) NMS. The NMS can be either not local to the switch, in which case another northbound interface is required for LSP attribute changes, or local/collocated, in which case there are significant issues with efficiency in resource usage. A stateful PCE adds a few features that:

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

## 5.7. Resource Defragmentation

In networks with link bundles, if LSPs are dynamically allocated and released over time, the resource becomes fragmented. The overall available resource on a (bundle) link might be sufficient for a new LSP request, but if the available resource is not continuous, the request is rejected. In order to perform the defragmentation procedure, stateful PCEs can be used, since global visibility of LSPs in the network is required to accurately assess resources on the

LSPs, and perform de-fragmentation while ensuring a minimal disruption of the network. This use case cannot be accommodated by a stateless PCE since it does not possess the detailed information of existing LSPs in the network.

A case of particular interest to GMPLS-based transport networks is the frequency defragmentation in flexible grid. In Flexible grid networks [I-D.ogrcetal-ccamp-flexi-grid-fwk], LSPs with different slot widths (such as 12.5G, 25G etc.) can co-exist so as to accommodate the services with different bandwidth requests. Therefore, even if the overall spectrum can meet the service request, it may not be usable if it is not contiguous. Thus, with the help of existing LSP state information, stateful PCE can make the resource grouped together to be usable. Moreover, stateful PCE can proactively choose routes for upcoming path requests to reduce the chance of spectrum fragmentation.

#### 5.8. Impairment-Aware Routing and Wavelength Assignment (IA-RWA)

In WSONs [RFC6163], a wavelength-switched LSP traverses one or more fiber links. The bit rates of the client signals carried by the wavelength LSPs may be the same or different. Hence, a fiber link may transmit a number of wavelength LSPs with equal or mixed bit rate signals. For example, a fiber link may multiplex the wavelengths with only 10G signals, mixed 10G and 40G signals, or mixed 40G and 100G signals.

IA-RWA in WSONs refers to the RWA process (i.e., lightpath computation) that takes into account the optical layer/transmission imperfections by considering as additional (i.e., physical layer) constraints. To be more specific, linear and non-linear effects associated with the optical network elements should be incorporated into the route and wavelength assignment procedure. For example, the physical imperfection can result in the interference of two adjacent lightpaths. Thus, a guard band should be reserved between them to alleviate these effects. The width of the guard band between two adjacent wavelengths depends on their characteristics, such as modulation formats and bit rates. Two adjacent wavelengths with different characteristics (e.g., different bit rates) may need a wider guard band and with same characteristics may need a narrower guard band. For example, 50GHz spacing may be acceptable for two adjacent wavelengths with 40G signals. But for two adjacent wavelengths with different bit rates (e.g., 10G and 40G), a larger spacing such as 300GHz spacing may be needed. Hence, the characteristics (states) of the existing wavelength LSPs should be considered for a new RWA request in WSON.

In summary, when stateful PCEs are used to perform the IA-RWA

procedure, they need to know the characteristics of the existing wavelength LSPs. The impairment information relating to existing and to-be-established LSPs can be obtained by nodes in WSON networks via external configuration or other means such as monitoring or estimation based on a vendor-specific impair model. However, WSON related routing protocols, i.e., [I-D.ietf-ccamp-wson-signal-compatibility-ospf] and [I-D.ietf-ccamp-gmpls-general-constraints-ospf-te], only advertise limited information (i.e., availability) of the existing wavelengths, without defining the supported client bit rates. It will incur substantial amount of control plane overhead if routing protocols are extended to support dissemination of the new information relevant for the IA-RWA process. In this scenario, stateful PCE(s) would be a more appropriate mechanism to solve this problem. Stateful PCE(s) can exploit impairment information of LSPs stored in LSP-DB to provide accurate RWA calculation.

## 6. Security Considerations

This document does not introduce any new security considerations beyond those discussed in [I-D.ietf-pce-stateful-pce].

The following topics will be discussed in a future version of this document: whether use of a stateful PCE makes the network more or less secure, and security use cases if any.

## 7. Contributing Authors

The following people all contributed significantly to this document and are listed below in alphabetical order:

Ramon Casellas  
CTTC - Centre Tecnologic de Telecomunicacions de Catalunya  
Av. Carl Friedrich Gauss n7  
Castelldefels, Barcelona 08860  
Spain  
Email: ramon.casellas@cttc.es

Edward Crabbe  
Google, Inc.  
1600 Amphitheatre Parkway  
Mountain View, CA 94043  
US  
Email: edc@google.com

Dhruv Dhody

Huawei Technology  
Leela Palace  
Bangalore, Karnataka 560008  
INDIA  
EMail: dhruvd@huawei.com

Oscar Gonzalez de Dios  
Telefonica Investigacion y Desarrollo  
Emilio Vargas 6  
Madrid, 28045  
Spain  
Phone: +34 913374013  
Email: ogondio@tid.es

Young Lee  
Huawei  
1700 Alma Drive, Suite 100  
Plano, TX 75075  
US  
Phone: +1 972 509 5599 x2240  
Fax: +1 469 229 5397  
EMail: ylee@huawei.com

Jan Medved  
Cisco Systems, Inc.  
170 West Tasman Dr.  
San Jose, CA 95134  
US  
Email: jmedved@cisco.com

Robert Varga  
Pantheon Technologies LLC  
Mlynske Nivy 56  
Bratislava 821 05  
Slovakia  
Email: robert.varga@pantheon.sk

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

Xiaobing Zi  
Email: unknown

## 8. Acknowledgements

We would like to thank Cyril Margaria, Adrian Farrel and JP Vasseur for the useful comments and discussions.

## 9. References

### 9.1. Normative References

- [I-D.ietf-pce-stateful-pce]  
Crabbe, E., Medved, J., Minei, I., and R. Varga, "PCEP Extensions for Stateful PCE",  
draft-ietf-pce-stateful-pce-04 (work in progress),  
May 2013.
- [RFC4655] Farrel, A., Vasseur, J., and J. Ash, "A Path Computation Element (PCE)-Based Architecture", RFC 4655, August 2006.
- [RFC5440] Vasseur, JP. and JL. Le Roux, "Path Computation Element (PCE) Communication Protocol (PCEP)", RFC 5440, March 2009.

### 9.2. Informative References

- [I-D.crabbe-pce-stateful-pce-mpls-te]  
Crabbe, E., Medved, J., Minei, I., and R. Varga, "Stateful PCE extensions for MPLS-TE LSPs",  
draft-crabbe-pce-stateful-pce-mpls-te-01 (work in progress), May 2013.
- [I-D.ietf-ccamp-gmpls-general-constraints-ospf-te]  
Zhang, F., Lee, Y., Han, J., Bernstein, G., and Y. Xu, "OSPF-TE Extensions for General Network Element Constraints",  
draft-ietf-ccamp-gmpls-general-constraints-ospf-te-04 (work in progress), July 2012.
- [I-D.ietf-ccamp-wson-signal-compatibility-ospf]  
Lee, Y. and G. Bernstein, "GMPLS OSPF Enhancement for Signal and Network Element Compatibility for Wavelength Switched Optical Networks",  
draft-ietf-ccamp-wson-signal-compatibility-ospf-11 (work in progress), February 2013.
- [I-D.ietf-pce-gmpls-pcep-extensions]  
Margaria, C., Dios, O., and F. Zhang, "PCEP extensions for GMPLS", draft-ietf-pce-gmpls-pcep-extensions-07 (work in progress), May 2013.



progress), October 2012.

- [I-D.ogrcetal-ccamp-flexi-grid-fwk]  
Dios, O., Casellas, R., Zhang, F., Fu, X., Ceccarelli, D.,  
and I. Hussain, "Framework and Requirements for GMPLS  
based control of Flexi-grid DWDM networks",  
draft-ogrcetal-ccamp-flexi-grid-fwk-02 (work in progress),  
February 2013.
- [I-D.sivabalan-pce-disco-stateful]  
Sivabalan, S., Medved, J., and X. Zhang, "IGP Extensions  
for Stateful PCE Discovery",  
draft-sivabalan-pce-disco-stateful-01 (work in progress),  
April 2013.
- [MPLS-PC] Chaieb, I., Le Roux, JL., and B. Cousin, "Improved MPLS-TE  
LSP Path Computation using Preemption", Global  
Information Infrastructure Symposium, July 2007.
- [MXMN-TE] Danna, E., Mandal, S., and A. Singh, "Practical linear  
programming algorithm for balancing the max-min fairness  
and throughput objectives in traffic engineering", pre-  
print, 2011.
- [NET-REC] Vasseur, JP., Pickavet, M., and P. Demeester, "Network  
Recovery: Protection and Restoration of Optical, SONET-  
SDH, IP, and MPLS", The Morgan Kaufmann Series in  
Networking, June 2004.
- [RFC3209] Awduche, D., Berger, L., Gan, D., Li, T., Srinivasan, V.,  
and G. Swallow, "RSVP-TE: Extensions to RSVP for LSP  
Tunnels", RFC 3209, December 2001.
- [RFC4427] Mannie, E. and D. Papadimitriou, "Recovery (Protection and  
Restoration) Terminology for Generalized Multi-Protocol  
Label Switching (GMPLS)", RFC 4427, March 2006.
- [RFC4657] Ash, J. and J. Le Roux, "Path Computation Element (PCE)  
Communication Protocol Generic Requirements", RFC 4657,  
September 2006.
- [RFC5212] Shiimoto, K., Papadimitriou, D., Le Roux, JL., Vigoureux,  
M., and D. Brungard, "Requirements for GMPLS-Based Multi-  
Region and Multi-Layer Networks (MRN/MLN)", RFC 5212,  
July 2008.
- [RFC5394] Bryskin, I., Papadimitriou, D., Berger, L., and J. Ash,  
"Policy-Enabled Path Computation Framework", RFC 5394,

December 2008.

- [RFC5521] Oki, E., Takeda, T., and A. Farrel, "Extensions to the Path Computation Element Communication Protocol (PCEP) for Route Exclusions", RFC 5521, April 2009.
- [RFC5557] Lee, Y., Le Roux, JL., King, D., and E. Oki, "Path Computation Element Communication Protocol (PCEP) Requirements and Protocol Extensions in Support of Global Concurrent Optimization", RFC 5557, July 2009.
- [RFC5623] Oki, E., Takeda, T., Le Roux, JL., and A. Farrel, "Framework for PCE-Based Inter-Layer MPLS and GMPLS Traffic Engineering", RFC 5623, September 2009.
- [RFC6163] Lee, Y., Bernstein, G., and W. Imajuku, "Framework for GMPLS and Path Computation Element (PCE) Control of Wavelength Switched Optical Networks (WSOs)", RFC 6163, April 2011.

#### Appendix A. Editorial notes and open issues

This section will be removed prior to publication.

The following open issues remain:

Use cases from draft-ietf-pce-stateful-pce To avoid loss of information, the use cases will be removed from [I-D.ietf-pce-stateful-pce] only after this document becomes a working group document.

This document WILL NOT repeat terminology defined in other documents or attempt to place any additional requirements on stateful PCE.

#### Authors' Addresses

Xian Zhang (editor)  
Huawei Technologies  
F3-5-B R&D Center, Huawei Base Bantian, Longgang District  
Shenzhen, Guangdong 518129  
P.R.China

Email: zhang.xian@huawei.com

Ina Minei (editor)  
Juniper Networks, Inc.  
1194 N. Mathilda Ave.  
Sunnyvale, CA 94089  
US

Email: [ina@juniper.net](mailto:ina@juniper.net)



Network Working Group  
Internet-Draft  
Intended status: Standard Track

Xian Zhang  
Gang Xie  
Dhruv Dhody  
Huawei

Expires: January 04, 2014

July 05, 2013

## LSP Synchronization for Stateful Path Computation Element (PCE)

draft-zhx-pce-stateful-lsp-sync-00.txt

### 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.

[Stateful-pcep] specifies a set of extensions to PCEP to enable stateful control of MPLS-TE and GMPLS Label Switched Paths (LSPs) via PCEP and maintaining of these LSPs at the stateful PCE. This document describes the mechanisms for incremental LSP Database (LSP-DB) synchronization as well as PCE control of the LSP-DB synchronization process.

### Status of this Memo

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

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

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

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

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

This Internet-Draft will expire on January 04, 2014.

## Copyright Notice

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

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

## Table of Contents

1. Introduction .....	2
1.1. Motivation .....	3
1.2. Conventions used in this document .....	4
2. PCEP Requirements and Objective .....	4
3. LSP Synchronization Procedure.....	4
3.1. New PCEP extensions .....	4
3.2. Procedure .....	4
4. IANA Considerations .....	6
5. Manageability Considerations .....	6
6. Security Considerations .....	6
7. Contributors .....	6
8. References .....	7
8.1. Normative References.....	7
Authors' Addresses .....	7

## 1. Introduction

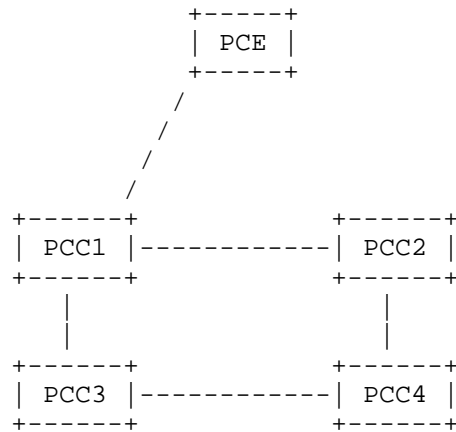
[Stateful-pcep] describes a Label Switched Path (LSP) state synchronization mechanism between Path Computation Clients (PCCs) and PCEs for a stateful PCE. It includes mechanisms for LSP state synchronization and avoidance between a PCC and a PCE when the PCEP session restarts. After PCEP session set up, PCC compares the LSP State Database version with the PCE. If the database version is mismatched, state synchronization will be performed. During state synchronization, a PCC sends the information of all its LSPs (full LSP-DB) to the stateful PCE.

This document proposes a mechanism for incremental (Delta) LSP Database (LSP-DB) synchronization as well as allowing PCE to control the timing of the LSP-DB synchronization process.

## 1.1. Motivation

If a PCE restarts and its LSP-DB survived, all PCCs with mismatched LSP State Database version will send all their LSPs information (full LSP-DB) to the stateful PCE, even if only a small number of LSPs underwent state change. It can take a long time and consume large communication channel bandwidth. Moreover, the stateful PCE can get overloaded with all the PCC performing full synchronization with it at the same time.

Figure 1 shows an example of LSP state synchronization.



Assuming there are 320 LSPs in the network, with each PCC having 80 LSPs. During the time when the PCEP session is down, 20 LSPs of each PCC (i.e., 80 LSPs in total), are changed. Hence when PCEP session restarts, the stateful PCE needs to synchronize 320 LSPs with all PCCs. But actually, 240 LSPs stay the same. If performing full LSP state synchronization, it can take a long time to carry out the synchronization of all LSPs. It is especially true when only a low bandwidth communication channel is available and there is a substantial number of LSPs in the network. Another disadvantage of full LSP synchronization is that it is a waste of communication bandwidth to perform full LSP synchronization given the fact that the number of LSP changes can be small during the time when PCEP session is down.

An incremental (Delta) LSP Database (LSP-DB) state synchronization is described in this document, where only the LSPs underwent state change are synchronized between the session restart. This may include new/modify/deleted LSPs. Furthermore, to avoid overloading the PCE, the proposed method enable a stateful PCE to control the LSP synchronization timing.

## 1.2. Conventions used in this document

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC-2119 [RFC2119].

## 2. PCEP Requirements and Objective

PCEP extensions for stateful PCEs to perform LSP synchronization SHOULD allow:

- Incremental LSP state synchronization between session restarts. Note this does not exclude the need for a stateful PCE to request a full LSP DB synchronization.
- A stateful PCE to control the timing of PCC synchronizing its LSP state with the PCE.

## 3. LSP Synchronization Procedure

[Stateful-pcep] describes state synchronization as well as state synchronization avoidance by using LSP-DB-VERSION TLV in its OPEN object. This document extends this idea to only synchronize the delta (changes) in case of version mismatch as well as to allow a stateful PCE to control the timing of this process.

### 3.1. New PCEP extensions

Two new bits are added in the STATEFUL-PCE-CAPABILITY TLV defined in [Stateful-pcep] for incremental (delta) LSP synchronization and PCE control:

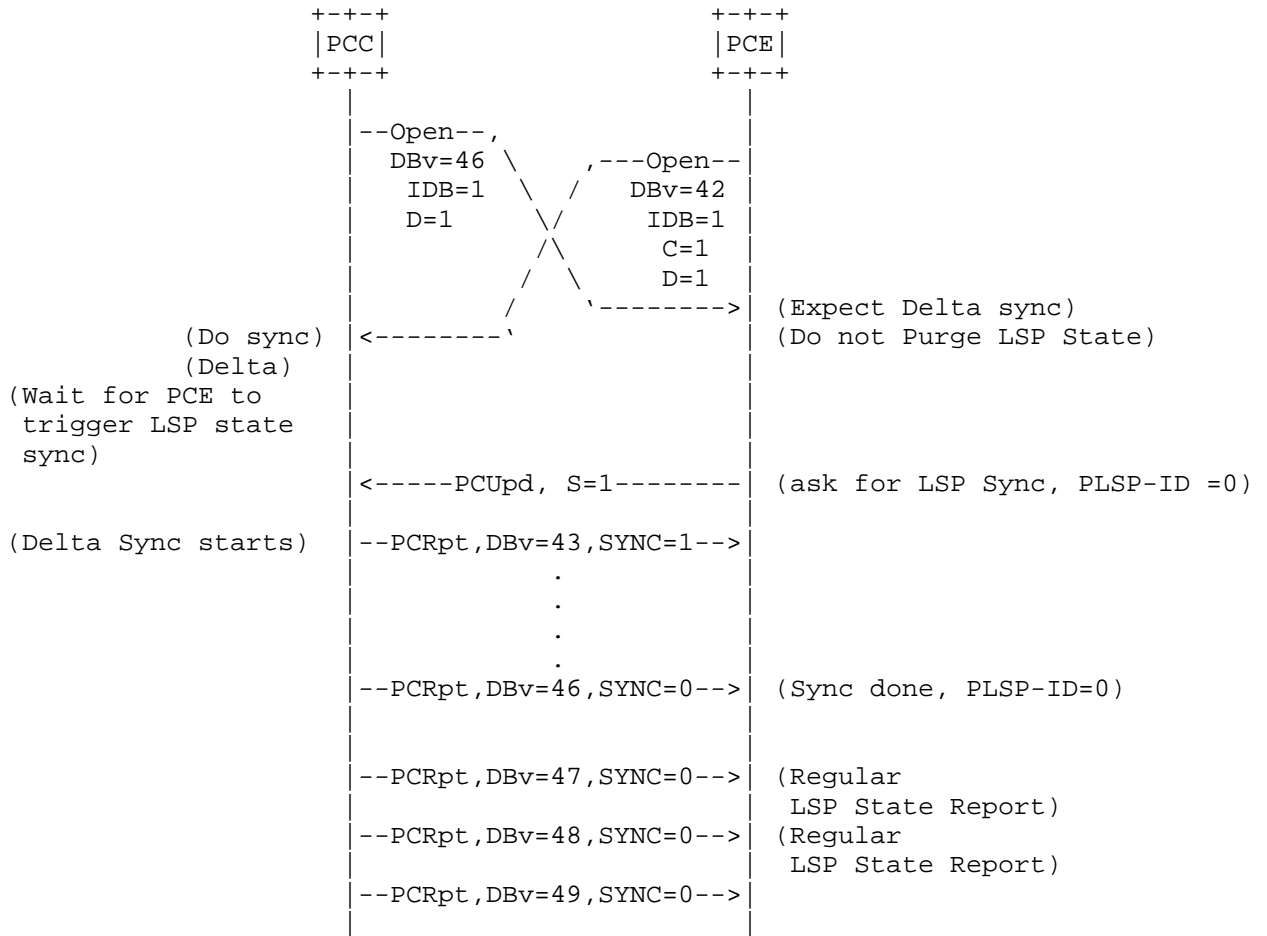
D (DELTA-LSP-SYNC-CAPABILITY - 1 bit): if set to 1 by a PCEP speaker, the D Flag indicates that the PCEP speaker allows delta or incremental state synchronization.

C (PCE-CONTROL-SYNC - 1 bit): if set to 1 by a stateful PCE, the C Flag indicates that the PCE will control the triggering of LSP state synchronization. This bit is not used by a PCC and MUST be set to 0 and ignored by the PCE upon receipt.

### 3.2. Procedure

If both PCEP speakers include the LSP-DB-VERSION TLV in the OPEN Object and the TLV values match, the PCC MAY skip state synchronization. Otherwise, the PCC MUST perform state synchronization. Instead of dumping full LSP-DB to PCE again, the PCC synchronizes the delta (changes) as described in figure 1 when D flag is set to 1 by both PCC and PCE. Other combinations of D flag setting by PCC and PCE result in full LSP-DB synchronization procedure as described in [Stateful-pcep].





A stateful PCE MAY choose to control the LSP-DB synchronization process. To allow PCE to do so, it MUST set C bit to 1 to indicate this. If the LSP DB version is mis-matched, it can send a PCUpd message with PLSP-ID = 0 and S =1 in order to trigger the LSP-DB synchronization process. In this way, the PCE can control the sequence of LSP synchronization among all the PCCs that re-establishing PCEP sessions with it. When the capability of PCE control is enable, only after a PCC receives this message, it will then start sending information that PCE does not possess, which is

inferred from the LSP DB Version information exchange in the OPEN message.

As per [Stateful-pcep], the LSP State Database version is incremented each time a change is made to the PCC's local LSP State Database. Each LSP is associated with the DB version at the time of its state change. This is needed to determine which LSP and what information needs to be synchronized in incremental state synchronization.

In the example shown in Figure 1, PCC synchronizes all LSPs that are updated between DB Version 43 to 46. A PCC SHOULD remember the deleted LSP as well, so that PCRpt message with deleted status can be sent to the stateful PCE.

#### 4. IANA Considerations

##### 4.1. STATEFUL-PCE-CAPABILITY TLV

As discussed in Section 3.1, two new STATEFUL-PCE-CAPABILITY TLV Flag Field has been defined. IANA has made the following allocation from the PCEP "STATEFUL-PCE-CAPABILITY TLV Flag Field" sub-registry:

Bit	Description	Reference
TBD	DELTA-LSP-SYNC-CAPABILITY	[This I.D.]
TBD	PCE-CONTROL-SYNC-CAPABILITY	[This I.D.]

#### 5. Manageability Considerations

The procedure defined in this document does not incur new manageability issues and the issues described in [Stateful-pcep] should be followed.

#### 6. Security Considerations

NONE

#### 7. Contributors

Young Lee

Huawei Technologies  
5360 Legacy Dr. Building 3  
Plano, TX 75024  
USA

Phone: (469) 277-5838  
Email: leeyoung@huawei.com

## 8. References

### 8.1. Normative References

[Stateful-pcep] Crabbe, E., Medved, J., Varga, R., Minei, I., "PCEP Extensions for Stateful PCE", draft-ietf-pce-stateful-pce, work in progress.

### Authors' Addresses

Xian Zhang

Huawei Technologies  
Research Area F3-1B,  
Huawei Industrial Base,  
Shenzhen, 518129, China

Phone: +86-755-28972645  
Email: zhang.xian@huawei.com

Gang Xie  
Huawei Technologies

Research Area F3  
Huawei Industrial Base,  
Shenzhen, 518129, China

Email xiegang09@huawei.com

Dhruv Dhody  
Huawei Technologies

Leela Palace  
Bangalore, Karnataka 560008  
INDIA  
Email: dhruv.dhody@huawei.com

