

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

Zafar Ali
Siva Sivabalan
Clarence Filsfils
Cisco Systems
Robert Varga
Pantheon Technologies
Victor Lopez
Oscar Gonzalez de Dios
Telefonica I+D
Xian Zhang
Huawei

October 21, 2013

Path Computation Element Communication Protocol (PCEP)
Extensions for remote-initiated GMPLS LSP Setup
draft-ali-pce-remote-initiated-gmpls-lsp-02.txt

Status of this Memo

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

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

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

This Internet-Draft will expire on April 20, 2014.

Copyright Notice

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

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

document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

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

Abstract

Draft [I-D. draft-crabbe-pce-pce-initiated-lsp] specifies procedures that can be used for creation and deletion of PCE-initiated LSPs in the active stateful PCE model. However, this specification focuses on MPLS networks, and does not cover remote instantiation of paths in GMPLS-controlled networks. This document complements [I-D. draft-crabbe-pce-pce-initiated-lsp] by addressing the requirements for remote-initiated GMPLS LSPs.

Conventions used in this document

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

Table of Contents

1. Introduction.....	3
2. Use Cases	3
2.1. Single-layer provisioning from active stateful PCE	3
2.2. Multi-layer networks	4
2.2.1. Higher-layer signaling trigger	4
2.3. NMS-VNTM cooperation model (separated flavor)	6
3. Requirements for Remote-Initiated GMPLS LSPs	7
4. PCEP Extensions for Remote-Initiated GMPLS LSPs	7
4.1. Generalized Endpoint in LSP Initiate Message	8
4.2. GENERALIZED-BANDWIDTH object in LSP Initiate Message	8
4.3. Protection Attributes in LSP Initiate Message	9
4.4. ERO in LSP Initiate Object	9
4.4.1. ERO with explicit label control	9
4.4.2. ERO with Path Keys	9
4.4.3. Switch Layer Object	10
4.5. LSP delegation and cleanup	10

5. Security Considerations	10
6. IANA Considerations	11
6.1. PCEP-Error Object	11
7. Acknowledgments	11
8. References	11
8.1. Normative References	11
8.2. Informative References	11

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 network.

[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. This enables realization of 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). This document complements [I-D. draft-crabbe-pce-pce-initiated-lsp] by addressing the requirements for remote-initiated GMPLS LSPs. These requirements are covered in Section 3 of this draft. The PCEP extensions for remote initiated GMPLS LSPs are specified in Section 4.

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 instantiate 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.

[See PDF version of the document for Figures]

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

L0 PCE obtains resources information via control plane collecting LSAs messages. The PCE computes the path and sends a message to the optical equipment with Explicit Route Object (ERO) information.

2.2. Multi-layer networks

This use case assumes there is a multi-layer network composed by routers and optical equipment. According to [RFC5623], there are four 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 the following we have selected two use cases to explain the requirements considered in this draft, but the document is applicable to all four options.

2.2.1. Higher-layer signaling trigger

Figure 2 depicts a multi-layer network scenario similar to the one 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.

[See PDF version of the document for Figures]

Figure 2. Use case higher-layer signaling trigger

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. Remote 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

Internet-Draft draft-ali-pce-remote-initiated-gmpls-lsp-02.txt

GMPLS tunnel using R1-O1-O2-R2 path (this is referred as GMPLS tunnel1 in the following). The remote initiated LSP using techniques specified in document is used for this purpose.

2.3. 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.

[See PDF version of the document for Figures]

Figure 3. Use case NMS-VNTM cooperation model

A new L3 path is requested from NMS (e.g., via 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, L3 PCE does not return a Path. Consequently, NMS sends a message to the VNTM to initiate 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 initiate the lower layer LSP via UNI signaling in the routers. Similarly, VNTM may talk with L0-PCE to set-up the path in the optical domain.

Requirements for the remote initiated GMPLS LSP from VNTM to the router are the same as discussed in the previous use case. The remote initiated LSP using techniques specified in document is used for this purpose.

3. Requirements for Remote-Initiated GMPLS 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].
- GMPLS LSP creation requires specification of data flow specific traffic parameters (also known as Tspec), which are technology specific.
- GMPLS also specifics 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 specifics protection context for the LSP, as defined in [RFC4872] and [RFC4873].

4. PCEP Extensions for Remote-Initiated GMPLS LSPs

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

4.1. Generalized Endpoint in LSP Initiate 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". PCInitiate 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 = 3 (Unknown Object), Error-value = 2(unknown object type).

4.2. GENERALIZED-BANDWIDTH object in LSP Initiate Message

LSP initiate 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:

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

4.3. Protection Attributes in LSP Initiate 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 Initiate 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 PCInitiate message can be used to specify protection attributes of the LSP being instantiated by the PCE.

4.4. ERO in LSP Initiate 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.

4.4.1. 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 PCInitiate message can be used to specify label along with ERO to PCC for the LSP being instantiated by the active stateful PCE.

4.4.2. 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 Initiate 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].

4.4.3. 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 PCInitiate message to specify the switching layer (or layers) of the LSP being remotely initiated.

4.5. LSP delegation and cleanup

LSP delegation and cleanup procedure specified in [I-D. draft-ietf-pcep-gmpls-ext] are equally applicable to GMPLS LSPs and this document does not modify the associated usage.

5. Security Considerations

To be added in future revision of this document.

Internet-Draft draft-ali-pce-remote-initiated-gmpls-lsp-02.txt

6. IANA Considerations

6.1. PCEP-Error Object

This document defines the following new Error-Value:

Error-Type	Error Value
------------	-------------

6	Error-value=TBA: LSP Request TLV missing
---	--

7. Acknowledgments

The authors would like to thank George Swallow and Jan Medved for their comments.

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.
- [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.

8.2. 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.

Internet-Draft draft-ali-pce-remote-initiated-gmpls-lsp-02.txt

- [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.
- [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.

Author's 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

Internet-Draft draft-ali-pce-remote-initiated-gmpls-lsp-02.txt

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

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

Zafar Ali
Siva Sivabalan
Clarence Filsfils
Cisco Systems
Robert Varga
Pantheon Technologies
Victor Lopez
Oscar Gonzalez de Dios
Telefonica I+D
Xian Zhang
Huawei
October 21, 2013

Path Computation Element Communication Protocol (PCEP)
Extensions for remote-initiated LSP Usage
draft-ali-pce-remote-initiated-lsp-usage-00.txt

Status of this Memo

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

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

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

This Internet-Draft will expire on April 20, 2014.

Copyright Notice

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

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

Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

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

Abstract

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 MPLS and GMPLS LSP Setup specifications do not address this requirement. This draft addresses the requirement to specify on how PCC should use the remote 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	3
2.1. Bandwidth-on-demand	3
3. Remote Initiated LSP Usage Requirement	5
4. PCEP extension for PCEP Initiated LSP Usage Specification	5
4.1. LSP_TUNNEL_INTERFACE_ID Object in LSP Initiate Message	5
4.2. Communicating LSP usage to Egress node	7
5. Security Considerations	7
6. IANA Considerations	7
6.1. END-POINT Object	7
7. Acknowledgments	7
8. References	7
8.1. Normative References	8
8.2. Informative References	8

1. Introduction

[I-D. draft-crabbe-pce-pce-initiated-lsp] and [I-D. draft-ali-pce-remote-initiated-gmpls-lsp] describe the setup and teardown of PCE-initiated MPLS and GMPLS 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, when an active stateful PCE is used for managing remote-initiated MPLS or GMPLS 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 3.

This draft 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 4. PCEP extensions specified in this document are equally applicable to PCEP initiated MPLS as well as GMPLS LSPs.

2. Use Cases

2.1. Bandwidth-on-demand

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).

[See PDF version of the document for Figures]

Figure 1. Bandwidth on demand use case

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. 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

Internet-Draft draft-ali-pce-remote-initiated-lsp-usage-00.txt

(tunnel IP address at R2). Section 4 of this draft addresses the requirement to specify on how PCC and egress node for signaling should use the remote initiated LSPs.

3. 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.

4. PCEP extension for PCEP Initiated LSP Usage Specification

The requirement to specify on how PCC should use the PCEP initiated LSPs in outlined in Section 2. This subsection specifies PCEP extension used to satisfy this requirement.

4.1. LSP_TUNNEL_INTERFACE_ID Object in LSP Initiate 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 Initiate (PCInitiate) 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).
- 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.
- 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.

4.2. Communicating LSP usage to Egress node

PCE does not need to send LSP Initiate message to egress node 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 use content of the LSP_TUNNEL_INTERFACE_ID TLV in LSP Initiate 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].

The egress node uses information contained in the LSP_TUNNEL_INTERFACE_ID object in RSVP-TE Path message 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.

5. Security Considerations

To be added in future revision of this document.

6. IANA Considerations

6.1. END-POINT Object

This document extends the LSP Initiate 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:

Name	Class value	Type
----	-----	----
LSP_TUNNEL_INTERFACE_ID	TBA by Iana (Suggested:40)	1

7. Acknowledgments

The authors would like to thank George Swallow and Jan Medved for their comments.

8. References

Internet-Draft draft-ali-pce-remote-initiated-lsp-usage-00.txt

8.1. Normative References

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

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

[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.

[I-D. draft-ali-pce-remote-initiated-gmpls-lsp] Z. Ali, S. Sivabalan, C. Filsfils, R. Varga, V. Lopez, O. Dios, X. Zhang, " Path Computation Element Communication Protocol (PCEP) Extensions for remote-initiated GMPLS LSP Setup", draft-ali-pce-remote-initiated-gmpls-lsp, work in progress.

8.2. Informative References

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

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
Expires January 2014

Internet-Draft draft-ali-pce-remote-initiated-lsp-usage-00.txt

Telefonica I+D
Email: ogondio@tid.es

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

Internet Engineering Task Force
Internet-Draft
Intended status: Standards Track
Expires: April 25, 2014

S. Alvarez
S. Sivabalan
Cisco Systems, Inc.
October 22, 2013

PCE Path Profiles
draft-alvarez-pce-path-profiles-00

Abstract

This document describes extensions to the Path Computation Element (PCE) Communication Protocol (PCEP) to signal path profiles. A stateful or stateless path computation element (PCE) can maintain an association between a set of path parameters and a profile. A PCC can use the path profile to initiate a path computation request without having to specify a detailed list of path parameters. In addition, a PCC can use the path profile to implement local policies associated with a path.

Status of This Memo

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

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

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

This Internet-Draft will expire on April 25, 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. Requirements Language	2
2. Path Profiles	2
3. Procedures	3
3.1. Capability Advertisement	3
3.2. PCC-Initiated Paths	3
3.2.1. Point-to-Point Paths	4
3.2.2. Point-to-Multipoint Paths	5
3.3. PCE-Initiated Paths	5
4. Object Extensions	6
4.1. OPEN Object	6
4.2. PATH-PROFILE Object	6
5. Error Codes for PATH-PROFILE Object	7
6. IANA Considerations	7
7. Security Considerations	7
8. References	7
8.1. Normative References	8
8.2. Informative References	8
Authors' Addresses	9

1. Introduction

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. Path Profiles

A path profile represents a list of path computation parameters or attributes that a PCE owns. The PCC learns that association through the interaction with the PCE. If the PCC initiates the path setup, it uses the path profile as part of the path computation request. In its simplest form, the request will only include mandatory path request objects and a PATH-PROFILE object defined in Section 4.2. The PCE uses the path profile to identify the appropriate path computation parameters. Then, it performs the path computation and sends a reply listing the detailed path parameters used. If the PCE initiates the path setup, the PCE signals the path parameters and includes a PATH-PROFILE object to indicate the path profile id associated with the path.

3. Procedures

3.1. Capability Advertisement

PCEP speakers advertise their capability to support path profiles during the session initialization phase. They include the PATH-PROFILE-CAPABILITY TLV defined in Section 4.1 as part of the OPEN object. A PCEP speaker can only signal a path profile if both speakers advertised this capability. A speaker **MUST** send a PCErr message with Error-Type=4 (Not supported object), Error-value=1 (Not supported object class) and close the session if it receives a message with a path profile id, it supports the extensions in this document and both speakers did not advertise this capability.

3.2. PCC-Initiated Paths

A PCC **MAY** include a PATH-PROFILE object when sending a PCReq message. The PCE uses the path profile id to select the parameters to fulfill the request. The means by which the PCC learns about a particular path profile id and decides to include it in a PCReq message are outside the scope of this document. Similarly, the means by which the PCE selects a set of parameters based on the profile id for a specific request are outside the scope of this document. The PCE may be stateful or stateless.

A PCE may receive a path computation request with an unknown or invalid path profile id. The PCE sends a PCErr message with Error-Type=[TBA] (PATH-PROFILE Error), Error-value=1 (Unknown path profile) if the path profile id is not known to the PCE. The PCE sends a PCErr message with Error-Type=[TBA] (PATH-PROFILE Error), Error-value=2 (Invalid path profile) if the PCE knows about the path profile id, but considers the request invalid. The profile may be invalid because of the path type, the PCEP session type or for the originating PCC.

The PCE will determine whether to consider any additional optional objects included in a PCReq message based on policy. As illustrated in Section 3.2.1 and Section 3.2.2, the PCC **MAY** include other optional objects along with a PATH-PROFILE object as part of a path computation request. The PCC will use the processing-rule (P) flag in the common object header to signal whether it considers those objects mandatory or optional when the PCE performs path computation. Those objects may overlap with the path parameters that the PCE associates with the path profile id.

PCE policy may place different kinds of restrictions on PCReq messages that include a PATH-PROFILE object and additional parameters. A PCE **MUST** send an error message if it receives a

request with optional objects signaled as mandatory (P flag = 1) for path computation and PCE policy does not allow such behavior from the originating PCC. In that case, the PCE sends a PCErr message with Error-Type=[TBA] (PATH-PROFILE Error), Error-value=3 (Unexpected mandatory object). If the objects are signaled as optional (P flag = 0) for path computation, the PCE will decide based on policy whether to consider them or not. When sending the PCRep message for the request, the PCE will use the ignore (I) flag in the common object header to indicate to the PCC whether an object was ignored.

3.2.1. Point-to-Point Paths

[RFC5440] defines the basic structure of a PCReq message for point-to-point paths. This documents extends the message format as follows:

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

where:

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

<request> ::= <RP>
              <END-POINTS>
              [<PATH-PROFILE>]
              [<path-computation>]
```

where:

<path-computation> is the list of optional objects used for path computation as defined initially in [RFC5440] and modified in subsequent PCEP extensions.

If present in a PCReq message, the PATH-PROFILE object MUST be the first optional object in the request portion of the message.

3.2.2. Point-to-Multipoint Paths

[RFC6006] defines the basic structure of a PCReq message for point-to-multipoint paths. This document extends the message format as follows:

TBD

3.3. PCE-Initiated Paths

A PCE MAY include a PATH-PROFILE object when sending a PCInitiate message as defined in [I-D.crabbe-pce-pce-initiated-lsp]. The PCC can use the path profile id to select local behavior to apply to the path. The means by which the PCE selects a profile id for a specific PCInitiate message are outside the scope of this document. Similarly, the means by which the PCC selects the local behavior to apply to a path based on a path profile id are outside the scope of this document.

[I-D.crabbe-pce-pce-initiated-lsp] defines the basic structure of a PCInitiate message. This document extends the message format 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> ::= (<PCE-initiated-lsp-instantiation> |
                                <PCE-initiated-lsp-deletion>)
```

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

```
<PCE-initiated-lsp-deletion> ::= <SRP>
                                   <LSP>
```

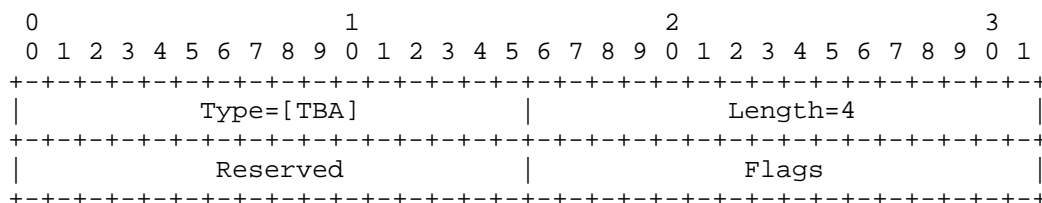
where:

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

4. Object Extensions

4.1. OPEN Object

This documents defines a new optional PATH-PROFILE-CAPABILITY TLV in the OPEN object.



PATH-PROFILE-CAPABILITY TLV

Figure 1

Reserved (16 bits):

MUST be set to zero on transmission and ignored on receipt.

Flags (16 bits):

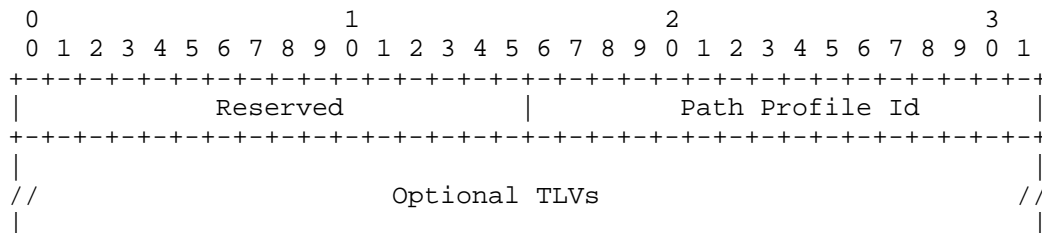
Unassigned bits are considered reserved. They MUST be set to zero on transmission and ignored on receipt. No flags are currently defined.

4.2. PATH-PROFILE Object

The PATH-PROFILE object may be carried in PCReq, PCInitiate and PCUpd messages.

PATH-PROFILE Object-Class is [TBA].

PATH-PROFILE Object-Type is 1.



+-----+

PATH-PROFILE Object

Figure 2

Reserved (16 bits):

MUST be set to zero on transmission and ignored on receipt.

Path Profile Id (16 bits):

(non-zero) unsigned path profile identifier.

The PATH-PROFILE object has a variable length and may contain additional TLVs. No TLVs are currently defined.

5. Error Codes for PATH-PROFILE Object

Error-Type	Meaning	Error-Value
<TBA>	PATH-PROFILE Error	1: Unknown path profile
		2: Invalid path profile
		3: Unexpected mandatory object

6. IANA Considerations

IANA is requested to assign the following code points.

PATH-PROFILE-CAPABILITY TLV

PATH-PROFILE Object-Class

PATH-PROFILE Object-Type

PATH-PROFILE Error-Type

7. Security Considerations

TBD

8. References

8.1. Normative References

- [I-D.ali-pce-remote-initiated-gmpls-lsp]
Ali, Z., Sivabalan, S., Filsfils, C., Varga, R., Lopez, V., and O. Dios, "Path Computation Element Communication Protocol (PCEP) Extensions for remote-initiated GMPLS LSP Setup", draft-ali-pce-remote-initiated-gmpls-lsp-01 (work in progress), July 2013.
- [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-03 (work in progress), October 2013.
- [I-D.ietf-pce-gmpls-pcep-extensions]
Margarita, C., Dios, O., and F. Zhang, "PCEP extensions for GMPLS", draft-ietf-pce-gmpls-pcep-extensions-08 (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-07 (work in progress), October 2013.
- [I-D.sivabalan-pce-segment-routing]
Sivabalan, S., Medved, J., Filsfils, C., Crabbe, E., and R. Raszuck, "PCEP Extensions for Segment Routing", draft-sivabalan-pce-segment-routing-02 (work in progress), October 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.
- [RFC6006] Zhao, Q., King, D., Verhaeghe, F., Takeda, T., Ali, Z., and J. Meuric, "Extensions to the Path Computation Element Communication Protocol (PCEP) for Point-to-Multipoint Traffic Engineering Label Switched Paths", RFC 6006, September 2010.

8.2. Informative References

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

Authors' Addresses

Santiago Alvarez
Cisco Systems, Inc.
170 West Tasman Drive
San Jose, CA 95134
USA

Email: saalvare@cisco.com

Siva Sivabalan
Cisco Systems, Inc.
2000 Innovation Drive
Kanata, ON K2K-3E8
Canada

Email: msiva@cisco.com

Internet Engineering Task Force
Internet-Draft
Intended status: Standards Track
Expires: April 24, 2014

H. Chen
Huawei Technologies
A. Liu
Ericsson
F. Xu
Verizon
M. Toy
Comcast
V. Liu
China Mobile
October 21, 2013

Extensions to PCEP for Distributing Label Cross Domains
draft-chen-pce-label-x-domains-00.txt

Abstract

This document specifies extensions to PCEP for distributing labels crossing domains for an inter-domain Point-to-Point (P2P) or Point-to-Multipoint (P2MP) Traffic Engineering (TE) Label Switched Path (LSP).

Status of this Memo

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

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

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

This Internet-Draft will expire on April 24, 2014.

Copyright Notice

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

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of

publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Table of Contents

1. Introduction	3
2. Terminology	3
3. Conventions Used in This Document	4
4. Label Distribution	4
4.1. An Exmaple	4
5. Extensions to PCEP	5
5.1. RP Object Extension	5
5.2. Label Object	6
5.3. LSP Tunnel Object	7
5.4. Request Message Extension	9
5.5. Reply Message Extension	9
6. Procedures	10
6.1. Distributing Label in Ordered Setup	10
6.2. Distributing Label in Path Computation	10
7. Security Considerations	11
8. IANA Considerations	11
8.1. Request Parameter Bit Flags	11
9. Acknowledgement	11
10. References	11
10.1. Normative References	11
10.2. Informative References	12
Authors' Addresses	12

1. Introduction

After a path crossing multiple domains is computed, an inter-domain Traffic Engineering (TE) Label Switched Path (LSP) tunnel may be set up along the path by a number of tunnel central controllers (TCCs). Each of the domains through which the path goes may be controlled by a tunnel central controller (TCC), which sets up the segment of the TE LSP tunnel in the domain. When the TCC sets up the segment of the TE LSP tunnel in its domain that is not a domain containing the tail end of the tunnel, it needs a label from a domain, which is next to it along the path.

This document specifies extensions to PCEP and various procedures for distributing a label from a domain to its previous domain along the path for the TE LSP tunnel crossing multiple domains.

2. Terminology

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

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

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

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

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

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

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

LSP: Label Switched Path.

LSR: Label Switching Router.

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

PCE: Path Computation Element. An entity (component, application, or

network node) that is capable of computing a network path or route based on a network graph and applying computational constraints.

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

TED: Traffic Engineering Database.

This document uses terminologies defined in RFC5440.

3. Conventions Used in This Document

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

4. Label Distribution

The Label Distribution may be provided by the PCE-based path computation. A PCE responsible for a domain computes a path segment for the domain, which is from an entry boundary to an exit boundary (or an egress) node of the domain. The PCE gets an label from the entry boundary node and adds an label object containing the label in the reply message to be sent to the requesting PCC (or another PCE).

When a PCE or PCC receives a reply message containing an label object, it removes the object from the message. The PCE may store the information in the label object or send the information to another component such as a Tunnel Central Controller (TCC).

4.1. An Exmaple

Figure 1 below illustrates a simple two-AS topology. There is a PCE responsible for the path computation in each AS. A path computation is requested from the Tunnel Central Controller (TCC), acting as the PCC, which sends the path computation request to PCE-1. PCE-1 is unable to compute an end-to-end path and invokes PCE-2 (possibly using the techniques described in [RFC5441]). PCE-2 computes a path segment from entry boundary node ASBR-2 of the right domain to the egress as {ASBR-2, C, D, Egress}. In addition to placing this path segment in the reply message to PCE-1, PCE-2 gets an label from the entry boundary node ASBR-2 and adds an label object containing the label and optionally the ASBR-2 into the reply message.

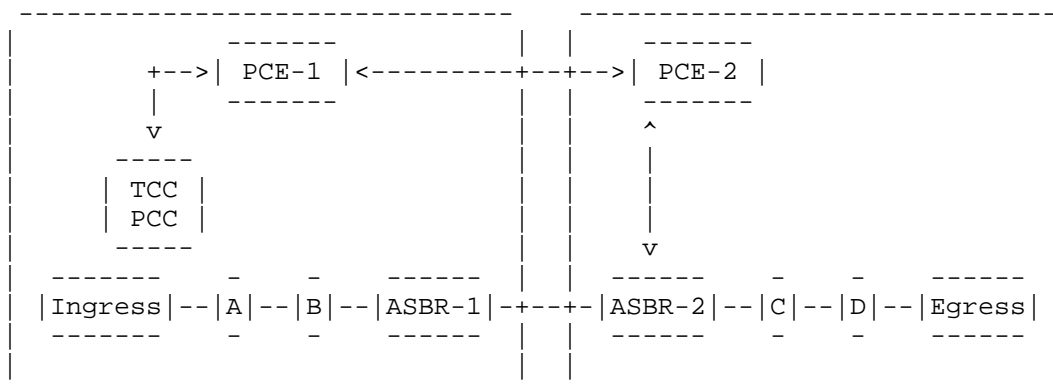


Figure 1: Example of Label Distribution

When PCE-1 receives the reply message containing the label object from PCE-2, it removes the object from the message. PCE-1 may store the information in the label object or send the information to another component such as a Tunnel Central Controller (TCC). TCC may set up the segment of the LSP tunnel from Ingress to ASBR-2 using the label in the label object from ASBR-2.

5. Extensions to PCEP

This section describes the extensions to PCEP for distributing labels crossing domains for an inter-domain Point-to-Point (P2P) or Point-to-Multipoint (P2MP) Traffic Engineering (TE) Label Switched Path (LSP). The extensions include the definition of a new flag in the RP object, tunnel information and label in a PCReq/PCRep message.

5.1. RP Object Extension

The following flags are added into the RP Object:

An L bit is added in the flag bits field of the RP object to tell a receiver of a PCReq/PCRep message that the message is for distributing labels crossing domains for an inter-domain LSP.

- o L (Label distribution bit - 1 bit):

- 0: This indicates that this is not a PCReq/PCRep message for distributing labels crossing domains.
- 1: This indicates that this is a PCReq or PCRep message for distributing labels crossing domains.

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

This L bit with the N bit defined in RFC6006 can indicate whether the PCReq/PCRep message is for distributing labels for an MPLS TE P2P LSP or an MPLS TE P2MP LSP.

- o L = 1 and N = 0: This indicates that this is a PCReq/PCRep message for distributing labels for a P2P LSP.
- o L = 1 and N = 1: This indicates that this is a PCReq/PCRep message for distributing labels for a P2MP LSP.

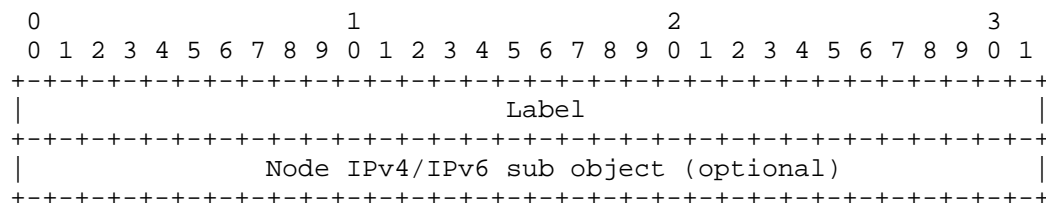
The C bit is added in the flag bits field of the RP object to tell the receiver of a PCReq/PCRep message that the message is for creating the segment of the LSP tunnel in a domain before distributing labels from this domain to its previous domain.

- o C (LSP tunnel Creation bit - 1 bit):
 - 0: This indicates that this is not a PCReq/PCRep message for creating the segment of the LSP tunnel.
 - 1: This indicates that this is a PCReq/PCRep message for creating the segment of the LSP tunnel in the domain before distributing labels to its previous domain.

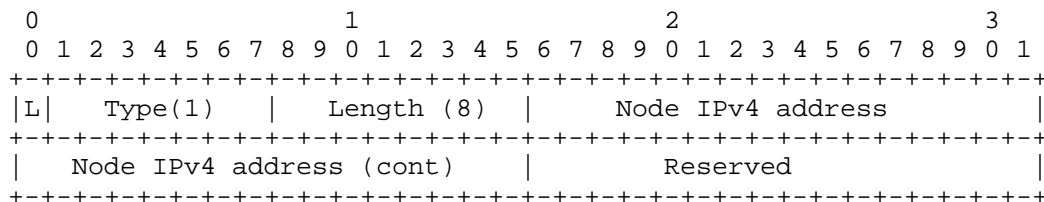
The IANA request is referenced in Section below (Request Parameter Bit Flags) of this document.

5.2. Label Object

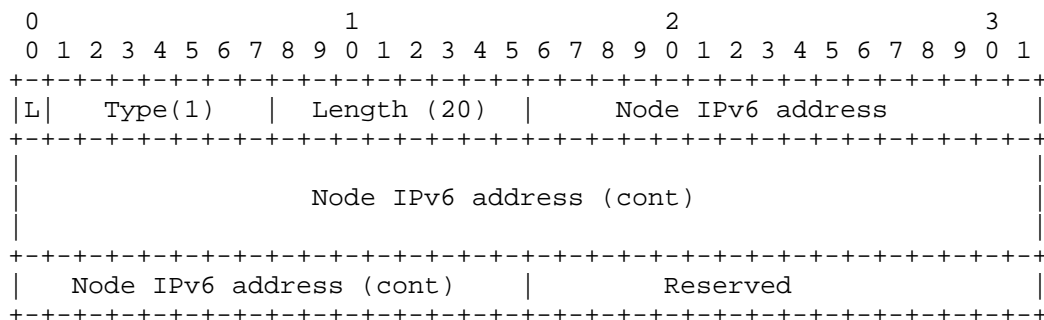
The format of a label object body (Object-Type=2) is illustrated below, which comprises a label and an optional node sub object. The node sub object contains a boundary node IP address, from which the label is allocated and distributed.



The format of the node IPv4 address sub object (Type=1) is as follows:



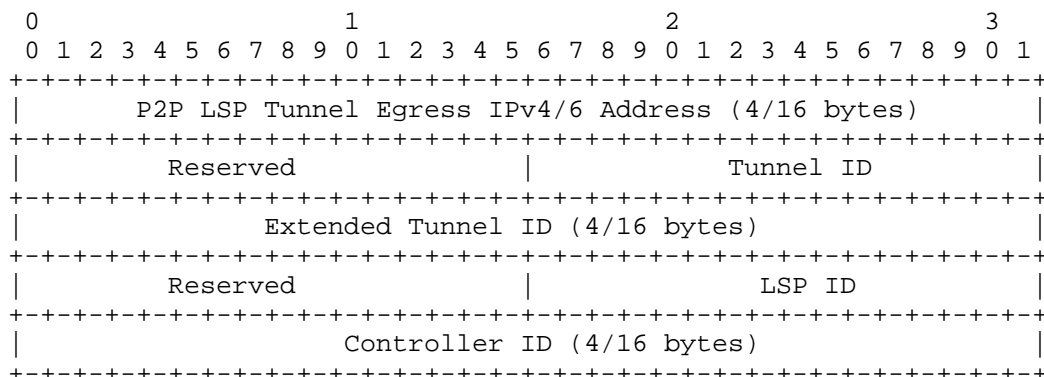
The format of the node IPv6 address sub object (Type=2) is illustrated below:



5.3. LSP Tunnel Object

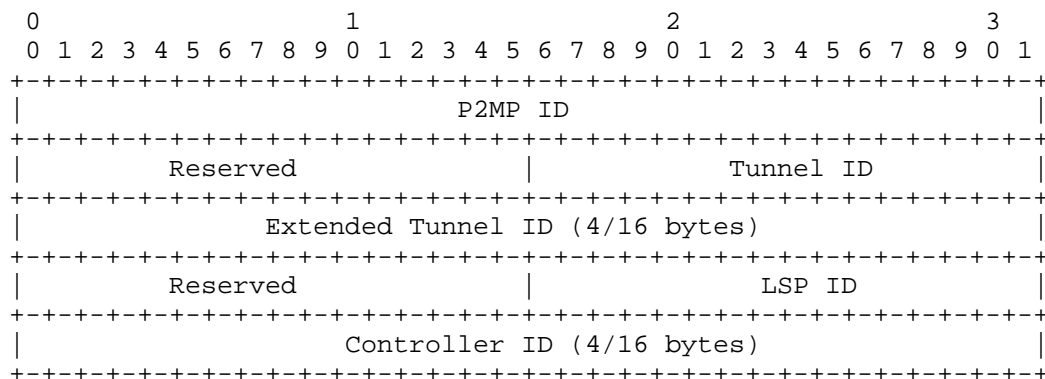
The LSP tunnel object contains the information that may be used to identify an LSP tunnel. An LSP tunnel may be a P2P or P2MP LSP tunnel. It may be an IPv4 or IPv6 LSP tunnel. Thus there are four types of LSP tunnels: 1) P2P LSP IPv4 tunnel, 2) P2P LSP IPv6 tunnel, 3) P2MP LSP IPv4 tunnel, and 4) P2MP LSP IPv6 tunnel.

The format of the P2P LSP IPv4/6 tunnel object body is as follows:



- o P2P LSP Tunnel Egress IPv4/6 Address:
IPv4/6 address of the egress of the tunnel.
- o Tunnel ID:
A 16-bit identifier that is constant over the life of the tunnel.
- o Extended Tunnel ID:
A 4/16-byte identifier that is constant over the life of the tunnel.
- o LSP ID:
A 16-bit identifier to allow resources sharing.
- o Controller ID:
A 4/16-byte identifier for the controller responsible for the head segment of the tunnel.

The format of the P2MP LSP IPv4/6 tunnel object body is as follows:



- o P2MP ID:
A 32-bit number unique within the ingress of LSP tunnel.
- o Tunnel ID:
A 16-bit identifier that is constant over the life of the tunnel.
- o Extended Tunnel ID:
A 4/16-byte identifier that is constant over the life of the tunnel.
- o LSP ID:
A 16-bit identifier to allow resources sharing.
- o Controller ID:
A 16-byte identifier for the controller responsible for the head segment of the tunnel.

5.4. Request Message Extension

Figure below illustrates the format of a request message with a optional LSP tunnel object:

```

<PCReq Message> ::= <Common Header>
                    [<svec-list>]
                    <request-list>
<request-list> ::= <request> [<request-list>]
<request> ::= <RP> <END-POINTS> [<OF>] [<LSPA>] [<BANDWIDTH>]
               [<metric-list>] [<RRO> [<BANDWIDTH>]] [<IRO>]
               [<LOAD-BALANCING>]
               [<LSP-tunnel>]

```

Figure 2: Format for Request Message

5.5. Reply Message Extension

Below is the format of a reply message with an optional Label object:

```

<PCReq Message> ::= <Common Header>
                    <response-list>
<response-list> ::= <response> [<response-list>]
<response> ::= <RP>
               [<NO-PATH>]
               [<attribute-list>]
               [<path-list>]
<path-list> ::= <path> [<path-list>]
<path> ::= <ERO> <attribute-list> [<LSP-tunnel>] [<Label>]

```

Figure 3: Format for Reply Message

6. Procedures

There may be a number of procedures for distributing labels crossing domains.

6.1. Distributing Label in Ordered Setup

Suppose that a path for an MPLS TE LSP tunnel crossing multiple domains is computed by PCEs and a sequence of domains (D_1, D_2, \dots, D_n) through which the path goes are controlled by a sequence of Tunnel Central Controllers TCCs ($TCC_1, TCC_2, \dots, TCC_n$) respectively. The method or procedure for distributing a label in ordered setup may comprise the following steps:

Step 1: TCC_i ($i = 1, \dots, n-1$) sends TCC_j ($j = i + 1$) a request for establishing the TE LSP tunnel.

Step 2: TCC_n (e.g., TCC_3) allocates a label from the enter border node (e.g., border node R) of domain D_n (e.g., D_3) and sends TCC_{n-1} (e.g., TCC_2) a reply containing the label after establishing the TE LSP tunnel segment (e.g., from node R to U) in domain D_n (e.g., D_3).

Step 3: TCC_j ($j = n-1, \dots, 2$) receives a reply containing a first label from TCC_{j+1} , allocates a second label from the enter border node of domain D_j , establishes the TE LSP tunnel segment in D_j and sends TCC_i ($i = j - 1$) a reply containing the label.

Step 4: TCC_1 receives a reply containing a label from TCC_2 and establishes the TE LSP tunnel segment in D_1 . At this point, the TE LSP tunnel crossing multiple domains is established.

6.2. Distributing Label in Path Computation

Suppose that a path for an MPLS TE LSP tunnel crossing multiple domains is computed by PCEs and a sequence of domains (D_1, D_2, \dots, D_n) through which the path goes are controlled by a sequence of PCEs ($PCE_1, PCE_2, \dots, PCE_n$) as TCCs respectively. The method or procedure for distributing a label in path computation may comprise the following steps:

Step 1: After PCE_n (e.g., PCE_3) receives a path request for computing the path and determines that the path segment of the path in domain D_n (e.g., D_3) is on the best path, it allocates a label from the enter border node (e.g., R) of domain D_n (e.g., D_3) on the path, establishes the TE LSP tunnel segment in domain D_n and sends PCE_{n-1} (e.g., PCE_2) a path reply containing the label.

Step 2: When PCE_j ($j = n-1, \dots, 2$) receives a path reply containing a first label from PCE_{j+1} and determines that the path segment of the path in domain D_j (e.g., D₂) is on the best path, it allocates a second label from the enter border node of domain D_j, establishes the TE LSP tunnel segment in D_j and sends PCE_i ($i = j - 1$) a path reply containing the second label.

Step 3: After PCE₁ receives a path reply containing a label from PCE₂ and determines the path segment in domain D₁, it establishes the TE LSP tunnel segment in D₁. At this point, the TE LSP tunnel crossing multiple domains is established.

7. Security Considerations

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

8. IANA Considerations

This section specifies requests for IANA allocation.

8.1. Request Parameter Bit Flags

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

Bit	Description	Reference
18	Label Distribution (L-bit)	This I-D
19	LSP tunnel Creation (C-bit)	This I-D

9. Acknowledgement

The author would like to thank people for their valuable comments on this draft.

10. References

10.1. Normative References

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

- [RFC3209] Awduche, D., Berger, L., Gan, D., Li, T., Srinivasan, V., and G. Swallow, "RSVP-TE: Extensions to RSVP for LSP Tunnels", RFC 3209, December 2001.
- [RFC5440] Vasseur, JP. and JL. Le Roux, "Path Computation Element (PCE) Communication Protocol (PCEP)", RFC 5440, March 2009.
- [RFC6006] Zhao, Q., King, D., Verhaeghe, F., Takeda, T., Ali, Z., and J. Meuric, "Extensions to the Path Computation Element Communication Protocol (PCEP) for Point-to-Multipoint Traffic Engineering Label Switched Paths", RFC 6006, September 2010.

10.2. Informative References

- [RFC4655] Farrel, A., Vasseur, J., and J. Ash, "A Path Computation Element (PCE)-Based Architecture", RFC 4655, August 2006.
- [RFC5862] Yasukawa, S. and A. Farrel, "Path Computation Clients (PCC) - Path Computation Element (PCE) Requirements for Point-to-Multipoint MPLS-TE", RFC 5862, June 2010.

Authors' Addresses

Huaimo Chen
Huawei Technologies
Boston, MA
US

Email: huaimo.chen@huawei.com

Autumn Liu
Ericsson
CA
USA

Email: autumn.liu@ericsson.com

Fengman Xu
Verizon
2400 N. Glenville Dr
Richardson, TX 75082
USA

Email: fengman.xu@verizon.com

Mehmet Toy
Comcast
1800 Bishops Gate Blvd.
Mount Laurel, NJ 08054
USA

Email: mehmet_toy@cable.comcast.com

Vic Liu
China Mobile
No.32 Xuanwumen West Street, Xicheng District
Beijing, 100053
China

Email: liuzhiheng@chinamobile.com

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

E. Crabbe
Google, Inc.
I. Minei
Juniper Networks, Inc.
S. Sivabalan
Cisco Systems, Inc.
R. Varga
Pantheon Technologies SRO
October 10, 2013

PCEP Extensions for PCE-initiated LSP Setup in a Stateful PCE Model
draft-crabbe-pce-pce-initiated-lsp-03

Abstract

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

The 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 April 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	7
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
5.3.1. The Create flag	11
5.4. LSP deletion	11
6. LSP delegation and cleanup	12
7. Implementation status	12
8. IANA considerations	13
8.1. PCEP Messages	13
8.2. LSP Object	13
8.3. PCEP-Error Object	14
9. Security Considerations	14
9.1. Malicious PCE	14
9.2. Malicious PCC	15
10. Acknowledgements	15
11. References	15
11.1. Normative References	15
11.2. Informative References	16
Authors' Addresses	17

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 or PCE 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 and MUST have a PLSP-ID of 0.

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 Delegate flag in the LSP object. In addition to the Delegate flag, the PCC also sets the Create flag in the LSP object (see Section 5.3.1), to indicate that the LSP was created as a result of a PCInitiate message. This PCRpt message MUST include the SRP object, with the SRP-id-number used in the SRP object of the PCInitiate message. 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 as described in [I-D.ietf-pce-stateful-pce]. 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. A PCE may return a delegation to the PCC in order to facilitate re-delegation of its LSPs to an alternate PCE.

To indicate a delete operation, the PCE MUST use the R flag in the SRP object in a PCUpd message. 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 in the LSP object for the removed state. See details in Section 5.4.

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:

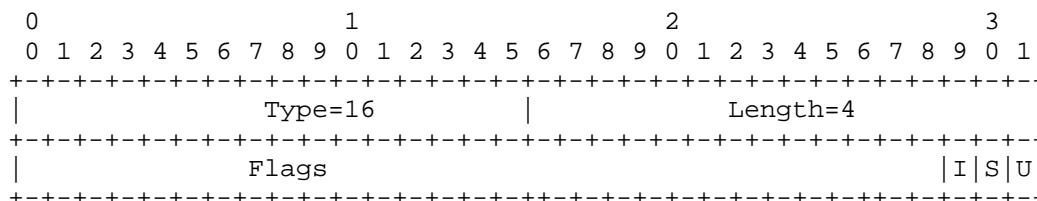


Figure 1: STATEFUL-PCE-CAPABILITY TLV format

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> ::= (<PCE-initiated-lsp-instantiation>|<PCE-initiated-lsp-deletion>)
```

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

```
<PCE-initiated-lsp-deletion> ::= <SRP>
                                   <LSP>
```

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:

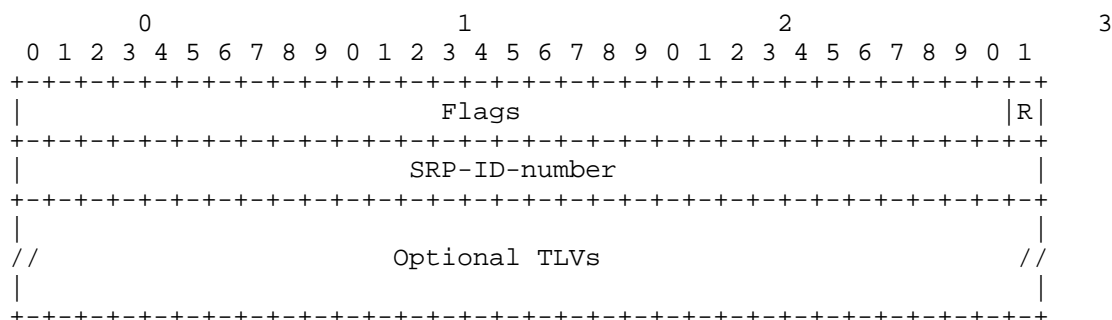


Figure 2: The SRP Object format

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. The LSP is set up using RSVP-TE, extensions for other setup methods are outside the scope of this draft.

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 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. LSPs that were instantiated as a result of a PCInitiate message MUST have the C flag set in the LSP object.

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). After the LSP is set up, errors in RSVP signaling are reported in PCRpt messages, as described in [I-D.ietf-pce-stateful-pce].

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.

Similarly, the PCE SHOULD be able to place a limit on either the number of LSP initiation requests pending for a particular PCC, or on the time it waits for a response (positive or negative) to a PCInitiate request from a PCC and MAY take further action (such as closing the session or removing all its LSPs) if this limit is

reached.

On successful completion of the LSP instantiation, the PCC assigns a PLSP-ID, and immediately delegates the LSP to the PCE by sending a PCRpt with the Delegate flag set. The PCRpt MUST include the SRP-ID-number of the PCInitiate request that triggered its creation. PCE-initiated LSPs are identified with the Create flag in the LSP Object.

5.3.1. The Create flag

The LSP object is defined in [I-D.ietf-pce-stateful-pce] and included here for easy reference.

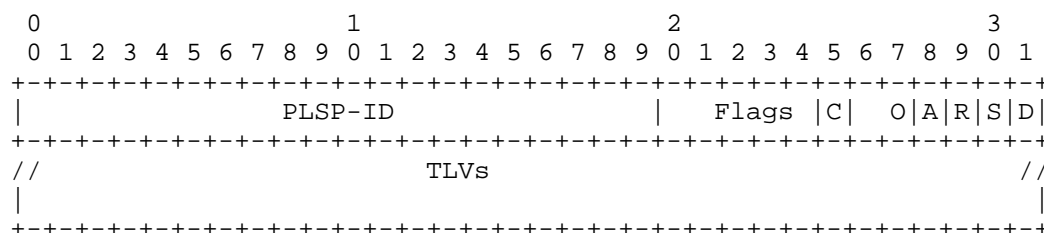


Figure 3: The LSP Object format

A new flag, the Create (C) flag is introduced. On a PCRpt message, the C Flag set to 1 indicates that this LSP was created via a PCInitiate message. The C Flag MUST be set to 1 on each PCRpt message for the duration of existence of the LSP. The Create flag allows PCEs to be aware of which LSPs were PCE-initiated (a state that would otherwise only be known by the PCC and the PCE that initiated them).

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])). If the PLSP-ID specified in the PCInitiate message was not created by the PCE, the PCC MUST send a PCErr message indicating "LSP is not PCE initiated" (Error code 19, error value TBD). Following the removal of the LSP, the PCC MUST send a PCRpt as described in [I-D.ietf-pce-stateful-pce]. The SRP object in the PCRpt MUST include the SRP-ID-number from the

PCInitiate message that triggered the removal. The R flag in the SRP object SHOULD be set.

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". The PCE MAY further react by closing the session.

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. After obtaining control of the LSP, the PCE may remove it using the procedures described in this document.

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. Implementation status

This section to be removed by the RFC editor.

This section records the status of known implementations of the protocol defined by this specification at the time of posting of this Internet-Draft, and is based on a proposal described in RFC 6982. The description of implementations in this section is intended to assist the IETF in its decision processes in progressing drafts to RFCs. Please note that the listing of any individual implementation here does not imply endorsement by the IETF. Furthermore, no effort has been spent to verify the information presented here that was supplied by IETF contributors. This is not intended as, and must not be construed to be, a catalog of available implementations or their features. Readers are advised to note that other implementations may exist.

According to RFC 6982, "this will allow reviewers and working groups to assign due consideration to documents that have the benefit of running code, which may serve as evidence of valuable experimentation and feedback that have made the implemented protocols more mature. It is up to the individual working groups to use this information as they see fit".

Two vendors are implementing the extensions described in this draft and have included the functionality in releases that will be shipping in the near future. An additional entity is working on implementing these extensions in the scope of research projects.

8. IANA considerations

8.1. PCEP Messages

This document defines the following new PCEP messages:

Value	Meaning	Reference
12	Initiate	This document

8.2. LSP Object

The following values are defined in this document for the Flags field in the LSP Object.

Bit	Description	Reference
24	Create	This document

8.3. 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

9. 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.

9.1. 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 for LSP instantiation 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.

9.2. Malicious PCC

The LSP instantiation mechanism described in this document requires the PCE to keep state for LSPs that it instantiates and relies on the PCC responding (with either a state report or an error message) to requests for LSP instantiation. A malicious PCC or one that reached the limit of the number of PCE-initiated LSPs, can ignore PCE requests and consume PCE resources. A PCE can protect itself by imposing a limit on the number of requests pending, or by setting a timeout and it MAY take further action such as closing the session or removing all the LSPs it initiated.

10. Acknowledgements

We would like to thank Jan Medved, Ambrose Kwong, Ramon Casellas, Dhruv Dhody, and Raveendra Trovi for their contributions to this document.

11. References

11.1. 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-07 (work in progress),
October 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, J.L., Vasseur, J.P., 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.

11.2. 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, 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.

[RFC6982] Sheffer, Y. and A. Farrel, "Improving Awareness of Running
Code: The Implementation Status Section", RFC 6982,
July 2013.

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: April 11, 2014

D. Dhody
F. Zhang
X. Zhang
Huawei Technologies
October 08, 2013

PCEP Extensions for Receiving SRLG Information
draft-dhody-pce-srlg-collection-00

Abstract

The Path Computation Element (PCE) provides functions of path computation in support of traffic engineering in networks controlled by Multi-Protocol Label Switching (MPLS) and Generalized MPLS (GMPLS).

This document provides extensions for the Path Computation Element Protocol (PCEP) to support collection of Shared Risk Link Group (SRLG) information during path computation and encoding this information in the reply message.

Status of This Memo

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

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

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

This Internet-Draft will expire on April 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	2
1.1. Requirements Language	3
2. Terminology	3
3. PCEP Requirements	3
4. Extension to PCEP	4
4.1. The Extension of the RP Object	4
4.2. SRLG Subobject in ERO	4
5. Other Considerations	5
5.1. Backward Compatibility	5
5.2. Confidentiality via PathKey	5
6. Security Considerations	5
7. Manageability Considerations	5
7.1. Control of Function and Policy	5
7.2. Information and Data Models	5
7.3. Liveness Detection and Monitoring	5
7.4. Verify Correct Operations	6
7.5. Requirements On Other Protocols	6
7.6. Impact On Network Operations	6
8. IANA Considerations	6
8.1. New Subobjects for the ERO Object	6
9. Acknowledgments	6
10. References	6
10.1. Normative References	6
10.2. Informative References	6
Appendix A. Contributor Addresses	8

1. Introduction

As per [RFC4655], PCE based path computation model is deployed in large, multi-domain, multi-region, or multi-layer networks. In such case PCEs may cooperate with each other to provide end to end optimal path.

It is important to understand which TE links in the network might be at risk from the same failures. In this sense, a set of links may constitute a 'shared risk link group' (SRLG) if they share a resource whose failure may affect all links in the set [RFC4202]. H-LSP (Hierarchical LSP) or S-LSP (Stitched LSP) can be used for carrying one or more other LSPs as described in [RFC4206] and [RFC6107]. H-LSP and S-LSP may be computed by PCE(s) and further form as a TE

link. The SRLG information of such LSPs can be collected during path computation itself and encoded in the PCEP Path Computation Reply (PCRep) message. [I-D.zhang-ccamp-gmpls-uni-app] describes the use of PCE for end to end User-Network Interface (UNI) path computation.

[I-D.farrel-interconnected-te-info-exchange] describes a scaling problem with SRLGs in multi-layer environment and introduce a concept of Macro SRLG. Lower layer SRLG collection at the time of path computation can be used to generate such a Macro SRLG at the PCE.

Note that [I-D.ietf-ccamp-rsvp-te-srlg-collect] specifies a similar extension to Resource ReserVation Protocol-Traffic Engineering (RSVP-TE) where SRLG information is collected at the time of signaling.

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.

CPS: Confidential Path Segment. A segment of a path that contains nodes and links that the policy requires not to be disclosed outside the domain.

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.

SRLG: Shared Risk Link Group.

UNI: User-Network Interface.

3. PCEP Requirements

Following key requirements are identified for PCEP to enable SRLG information collection during path computation:

SRLG Collection Indication: The PCEP speaker must be capable of indicating whether the SRLG information of the LSP should be collected during the path computation procedure.

SRLG Collection: If requested, the SRLG information should be collected during the path computation and encoded in the PCRep message.

4. Extension to PCEP

This document extends the existing RP (Request Parameters) object [RFC5440] so that a PCEP speaker can request SRLG information collection during path computation. The SRLG subobject maybe carried inside the Explicit Route Object (ERO) in the PCRep message.

4.1. The Extension of the RP Object

This document adds the following flags to the RP Object:

S (SRLG - 1 bit): when set, in a PCReq message, this indicates that the SRLG information of the Label switched path (LSP) should be collected during the path computation procedure. Otherwise, when cleared, this indicates that the SRLG information should not be collected. In a PCRep message, when the S bit is set this indicates that the returned path in ERO also carry the SRLG information; otherwise (when the S bit is cleared), the returned path does not carry SRLG information.

4.2. SRLG Subobject in ERO

As per [RFC5440], ERO is used to encode the path of a TE LSP and is carried within a PCRep message to provide the computed path when computation was successful.

The SRLG of a path is the union of the SRLGs of the links in the LSP [RFC4202]. The SRLG subobject is defined in [I-D.ietf-ccamp-rsvp-te-srlg-collect], as shown below:

```

      0                               1                               2                               3
      0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-----+-----+-----+-----+-----+-----+-----+-----+
|          Type          |      Length      |   Reserved   |     Flags     |
+-----+-----+-----+-----+-----+-----+-----+-----+
|          SRLG ID 1 (4 bytes)          |
+-----+-----+-----+-----+-----+-----+-----+-----+
~                               .....                               ~
+-----+-----+-----+-----+-----+-----+-----+-----+
|          SRLG ID n (4 bytes)          |
+-----+-----+-----+-----+-----+-----+-----+-----+

```

The meaning and description of Type, Length and SRLG ID can be found in [I-D.ietf-ccamp-rsvp-te-srlg-collect]. Bits in the Flags field is ignored.

The SRLG subobject should be encoded inside the ERO object in the PCRep message when the S-Bit (SRLG) is set in the PCReq message.

5. Other Considerations

5.1. Backward Compatibility

If a PCE receives a request and the PCE does not understand the new SRLG flag in the RP object, then the PCE SHOULD reject the request.

If PCEP speaker receives a PCRep message with SRLG subobject that it does not support or recognize, it must act according to the existing processing rules.

5.2. Confidentiality via PathKey

[RFC5520] defines a mechanism to hide the contents of a segment of a path, called the Confidential Path Segment (CPS). The CPS may be replaced by a path-key that can be conveyed in the PCEP and signaled within in a RSVP-TE ERO.

When path-key confidentiality is used, collection of SRLG information and encoding this information in PCRep along with the path-key could be useful to compute a SRLG disjoint backup path at the later instance.

6. Security Considerations

TBD.

7. Manageability Considerations

7.1. Control of Function and Policy

TBD.

7.2. Information and Data Models

TBD.

7.3. Liveness Detection and Monitoring

TBD.

7.4. Verify Correct Operations

TBD.

7.5. Requirements On Other Protocols

TBD.

7.6. Impact On Network Operations

TBD.

8. IANA Considerations

IANA assigns values to PCEP parameters in registries defined in [RFC5440]. IANA is requested to make the following additional assignments.

8.1. New Subobjects for the ERO Object

IANA has previously assigned an Object-Class and Object-Type to the ERO carried in PCEP messages [RFC5440]. IANA also maintains a list of subobject types valid for inclusion in the ERO.

IANA is requested to assign one new subobject types for inclusion in the ERO as follows:

Subobject Meaning	Reference
34 (TBD) SRLG sub-object	This document

9. Acknowledgments

TBD.

10. References

10.1. Normative References

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

10.2. Informative References

[RFC4202] Kompella, K. and Y. Rekhter, "Routing Extensions in Support of Generalized Multi-Protocol Label Switching (GMPLS)", RFC 4202, October 2005.

- [RFC4206] Kompella, K. and Y. Rekhter, "Label Switched Paths (LSP) Hierarchy with Generalized Multi-Protocol Label Switching (GMPLS) Traffic Engineering (TE)", RFC 4206, October 2005.
- [RFC4655] Farrel, A., Vasseur, J., and J. Ash, "A Path Computation Element (PCE)-Based Architecture", RFC 4655, August 2006.
- [RFC4874] Lee, CY., Farrel, A., and S. De Chodder, "Exclude Routes - Extension to Resource ReserVation Protocol-Traffic Engineering (RSVP-TE)", RFC 4874, April 2007.
- [RFC5440] Vasseur, JP. and JL. Le Roux, "Path Computation Element (PCE) Communication Protocol (PCEP)", RFC 5440, March 2009.
- [RFC5520] Bradford, R., Vasseur, JP., and A. Farrel, "Preserving Topology Confidentiality in Inter-Domain Path Computation Using a Path-Key-Based Mechanism", RFC 5520, April 2009.
- [RFC6107] Shiimoto, K. and A. Farrel, "Procedures for Dynamically Signaled Hierarchical Label Switched Paths", RFC 6107, February 2011.
- [I-D.ietf-ccamp-rsvp-te-srlg-collect]
Zhang, F., Li, D., Dios, O., Margaria, C., and M. Hartley, "RSVP-TE Extensions for Collecting SRLG Information", draft-ietf-ccamp-rsvp-te-srlg-collect-03 (work in progress), September 2013.
- [I-D.farrel-interconnected-te-info-exchange]
Farrel, A., Drake, J., Bitar, N., Swallow, G., and D. Ceccarelli, "Problem Statement and Architecture for Information Exchange Between Interconnected Traffic Engineered Networks", draft-farrel-interconnected-te-info-exchange-01 (work in progress), July 2013.
- [I-D.zhang-ccamp-gmpls-uni-app]
Zhang, F., Dios, O., Farrel, A., Zhang, X., and D. Ceccarelli, "Applicability of Generalized Multiprotocol Label Switching (GMPLS) User-Network Interface (UNI)", draft-zhang-ccamp-gmpls-uni-app-04 (work in progress), July 2013.

Appendix A. Contributor Addresses

Udayasree Palle
Huawei Technologies
Leela Palace
Bangalore, Karnataka 560008
INDIA

EMail: udayasree.palle@huawei.com

Avantika
Huawei Technologies
Leela Palace
Bangalore, Karnataka 560008
INDIA

EMail: avantika.sushilkumar@huawei.com

Authors' Addresses

Dhruv Dhody
Huawei Technologies
Leela Palace
Bangalore, Karnataka 560008
INDIA

EMail: dhruv.ietf@gmail.com

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

EMail: zhangfatai@huawei.com

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

EMail: zhang.xian@huawei.com

PCE Working Group
Internet-Draft
Intended status: Experimental
Expires: March 28, 2014

D. Dhody
Q. Wu
U. Pallé
X. Zhang
Huawei Technologies
September 24, 2013

PCE support for Domain Diversity
draft-dwpz-pce-domain-diverse-00

Abstract

The Path Computation Element (PCE) may be used for computing path for services that traverse multi-area and multi-AS Multiprotocol Label Switching (MPLS) and Generalized MPLS (GMPLS) Traffic Engineered (TE) networks.

Path computation should facilitate the selection of paths with domain diversity. This document examines the existing mechanisms to do so and further propose some extensions to Path Computation Element Protocol (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 March 28, 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. Requirements Language	3
2. Terminology	3
3. Domain Diversity	3
3.1. Per Domain Path Computation	4
3.2. Backward-Recursive PCE-based Computation	4
3.3. Hierarchical PCE	4
3.3.1. End to End Path	5
3.3.2. Domain-Sequence	5
4. Extension to PCEP	5
4.1. SVEC Object	5
4.2. Transit Domain Identifier	6
4.3. Minimize Shared Domains	6
5. Security Considerations	7
6. Manageability Considerations	7
6.1. Control of Function and Policy	7
6.2. Information and Data Models	7
6.3. Liveness Detection and Monitoring	7
6.4. Verify Correct Operations	7
6.5. Requirements On Other Protocols	7
6.6. Impact On Network Operations	7
7. IANA Considerations	7
8. Acknowledgments	8
9. References	8
9.1. Normative References	8
9.2. Informative References	8
Appendix A. Contributor Addresses	9

1. Introduction

The ability to compute shortest constrained TE LSPs in Multiprotocol Label Switching (MPLS) and Generalized MPLS (GMPLS) networks across multiple domains has been identified as a key requirement. In this context, a domain is a collection of network elements within a common sphere of address management or path computational responsibility such as an Interior Gateway Protocol (IGP) area or an Autonomous Systems (AS).

In a multi-domain environment, Domain Diversity is defined in [RFC6805]. A pair of paths are domain-diverse if they do not traverse any of the same transit domains. Domain diversity may be maximized for a pair of paths by selecting paths that have the smallest number of shared domains. Path computation should facilitate the selection of domain diverse paths as a way to reduce the risk of shared failure and automatically helps to ensure path diversity for most of the route of a pair of LSPs.

This document examine a way to achieve domain diversity with existing inter-domain path computation mechanism like per-domain path computation technique [RFC5152], Backward Recursive Path Computation (BRPC) mechanism [RFC5441] and Hierarchical PCE [RFC6805]. This document also considers synchronized dependent path computations as well as non-synchronized path computation. Since independent and synchronized path computation cannot be used to apply diversity, it is not discussed in this document.

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].

3. Domain Diversity

As described in [RFC6805], a set of paths are considered to be domain diverse if they do not share any transit domains, apart from ingress and egress domains.

Some additional parameters to consider would be -

Minimize shared domain: When a fully domain diverse path is not possible, PCE could be requested to minimize the number of shared transit domains. This can also be termed as maximizing partial domain diversity.

Boundary Nodes: TBD

3.1. Per Domain Path Computation

The per domain path computation technique [RFC5152] defines a method where the path is computed during the signaling process (on a per-domain basis). The entry Boundary Node (BN) of each domain is responsible for performing the path computation for the section of the LSP that crosses the domain, or for requesting that a PCE for that domain computes that piece of the path.

Non-Synchronized Path Computation: Path computations are performed in a serialized and independent fashion. After the setup of primary path, a domain diverse path can be signaled by encoding the transit domain identifiers in XRO or EXRS using domain sub-objects defined in [DOMAIN-SUBOBJ] and [RFC3209] in RSVP-TE. Note that the head end LSR should be aware of transit domain identifiers of the primary path to be able to do so.

Synchronized Path Computation: Not Applicable.

3.2. Backward-Recursive PCE-based Computation

The BRPC [RFC5441] technique involves cooperation and communication between PCEs in order to compute an optimal end-to-end path across multiple domains. The sequence of domains to be traversed maybe known before the path computation, but it can also be used when the domain path is unknown and determined during path computation.

Non-Synchronized Path Computation: Path computations are performed in a serialized and independent fashion. After the path computation and setup of primary path, a domain diverse path computation request is sent by PCC to the PCE, by encoding the transit domain identifiers in XRO or EXRS using domain sub-objects defined in [PCE-DOMAIN] and [RFC3209] in PCEP. Note that the PCC should be aware of transit domain identifiers of the primary path to be able to do so.

Synchronized Path Computation: Not Applicable. [Since different transit domain PCEs are involved , there is no way to achieve synchronization for domain diverse paths]. BTW [RFC5440] describes other diversity parameters (node, link etc).

3.3. Hierarchical PCE

In H-PCE [RFC6805] architecture, the parent PCE is used to compute a multi-domain path based on the domain connectivity information. The parent PCE may be requested to provide a end to end path or only the sequence of domains.

3.3.1. End to End Path

Non-Synchronized Path Computation: Path computations are performed in a serialized and independent fashion. After the path computation and setup of primary path, a domain diverse path computation request is sent to the parent PCE, by encoding the transit domain identifiers in XRO or EXRS using domain sub-objects defined in [PCE-DOMAIN] and [RFC3209] in PCEP. Note that the PCC should be aware of transit domain identifiers of the primary path to be able to do so. The parent PCE should provide a domain diverse end to end path.

Synchronized Path Computation: Child PCE should be able to request dependent and synchronized domain diverse end to end paths from its parent PCE. A new flag is added in SVEC object for this (Refer Section 4.1).

3.3.2. Domain-Sequence

Non-Synchronized Path Computation: Path computations are performed in a serialized and independent fashion. After the primary path computation using H-PCE (involving domain-sequence selection by parent PCE and end-to-end path computation via BRPC or Per-Domain mechanisms) and setup, a domain diverse path computation request is sent to the parent PCE, by encoding the transit domain identifiers in XRO or EXRS using domain sub-objects defined in [PCE-DOMAIN] and [RFC3209] in PCEP. Note that the PCC should be aware of transit domain identifiers of the primary path to be able to do so. The parent PCE should provide a diverse domain sequence.

Synchronized Path Computation: Child PCE should be able to request dependent and synchronized diverse domain-sequence(s) from its parent PCE. A new flag is added in SVEC object for this (Refer Section 4.1). The parent PCE should reply with diverse domain sequence(s) encoded in ERO as described in [PCE-DOMAIN].

4. Extension to PCEP

4.1. SVEC Object

[RFC5440] defines SVEC object which includes flags for the potential dependency between the set of path computation requests (Link, Node and SRLG diverse). This document proposes a new flag 0 for domain diversity.

The format of the SVEC object body is as follows:

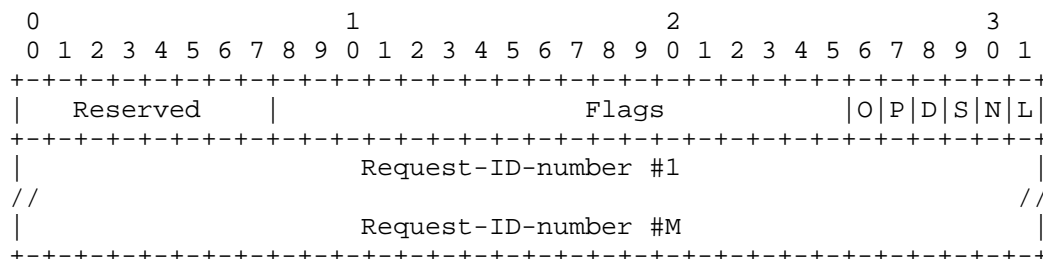


Figure 1: SVEC Body Object Format

Following new bit is added in the Flags field:

- * O (Domain diverse) bit: when set, this indicates that the computed paths corresponding to the requests specified by the following RP objects MUST NOT have any transit domain(s) in common.

The Domain Diverse O-bit can be used in Hierarchical PCE path computation to compute synchronized domain diverse end to end path or diverse domain sequences as described in Section 3.3.

When domain diverse O bit is set, it is applied to the transit domains. The other bit in SVEC object (N, L etc) is set, should still be applied in the ingress and egress domain.

4.2. Transit Domain Identifier

In case of non-synchronized path computation, Ingress node (i.e. a PCC) should be aware of transit domain identifiers of the primary path. So during the path computation or signaling of the primary path, the transit domain should be identified.

A possible solution for path computation could be a flag in RP object requesting domain identifier to be returned in the PCEP path reply message. Further details - TBD

4.3. Minimize Shared Domains

A new Objective function (OF) [RFC5541] code for synchronized path computation requests is proposed:

MCTD

- * Name: Minimize the number of Common Transit Domains.
- * Objective Function Code: TBD

- * Description: Find a set of paths such that it passes through the least number of common transit domains.

The MCTD OF can be used in Hierarchical PCE path computation to request synchronized domain diverse end to end paths or diverse domain sequences as described in Section 3.3.

For non synchronized diverse domain path computation the X bit in XRO or EXRS [RFC5521] sub-objects can be used, where X bit set as 1 indicates that the domain specified SHOULD be excluded from the path computed by the PCE, but MAY be included subject to PCE policy and the absence of a viable path that meets the other constraints and excludes the domain.

5. Security Considerations

TBD.

6. Manageability Considerations

6.1. Control of Function and Policy

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

TBD.

8. Acknowledgments

We would like to thank Qilei Wang for starting this discussion in the mailing list.

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

- [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.
- [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.
- [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.
- [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.
- [DOMAIN-SUBOBJ] Dhody, D., Palle, U., Kondreddy, V., and R. Casellas, "Domain Subobjects for Resource ReserVation Protocol -

Traffic Engineering (RSVP-TE). (draft-dhody-ccamp-rsvp-te-domain-subobjects)", July 2013.

[PCE-DOMAIN]

Dhody, D., Palle, U., and R. Casellas, "Standard Representation Of Domain Sequence. (draft-ietf-pce-pcep-domain-sequence)", July 2013.

Appendix A. Contributor Addresses

Ramon Casellas
CTTC - Centre Tecnologic de Telecomunicacions de Catalunya
Av. Carl Friedrich Gauss n7
Castelldefels, Barcelona 08860
SPAIN

EMail: ramon.casellas@cttc.es

Avantika
Huawei Technologies
Leela Palace
Bangalore, Karnataka 560008
INDIA

EMail: avantika.sushilkumar@huawei.com

Authors' Addresses

Dhruv Dhody
Huawei Technologies
Leela Palace
Bangalore, Karnataka 560008
INDIA

EMail: dhruv.ietf@gmail.com

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

EMail: bill.wu@huawei.com

Udayasree Palle
Huawei Technologies
Leela Palace
Bangalore, Karnataka 560008
INDIA

EMail: udayasree.palle@huawei.com

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

EMail: zhang.xian@huawei.com

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

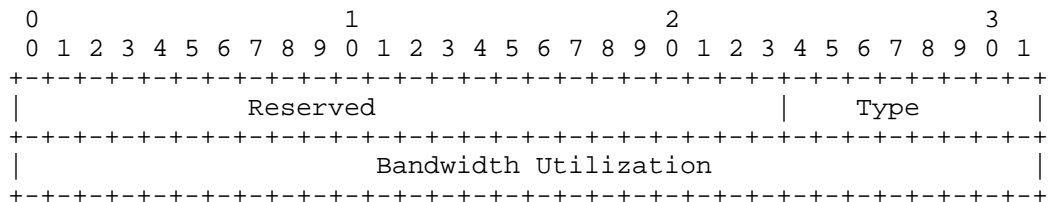
The BU object is used to indicate the upper limit of the acceptable link bandwidth utilization percentage.

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:



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 $(\text{Min} \{ (M(L_{pi}) - u(L_{pi})) / M(L_{pi}), i=1 \dots K \})$ is maximized.

Objective Function Code: TBD7

Name: Maximum Reserved Under-Utilized Path (MRUP)

Description: Find a path P such that $(\text{Min} \{ (R(L_{pi}) - ru(L_{pi})) / R(L_{pi}), i=1 \dots K \})$ 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.

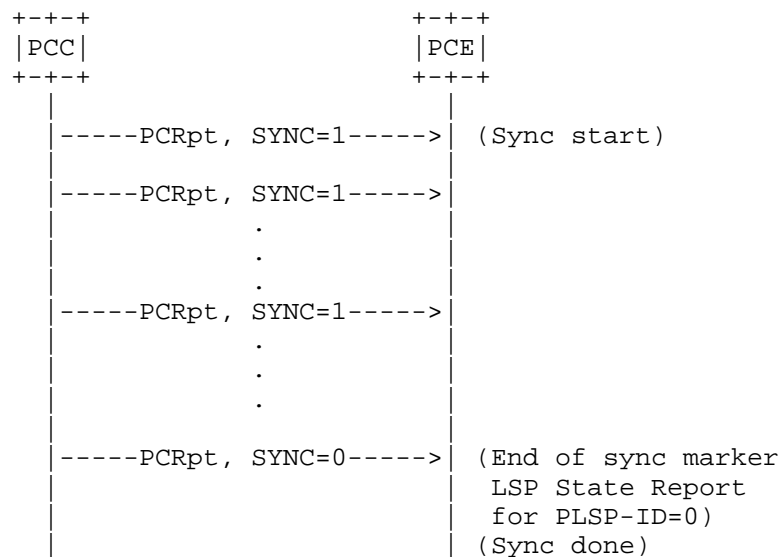


Figure 1: Successful state synchronization

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

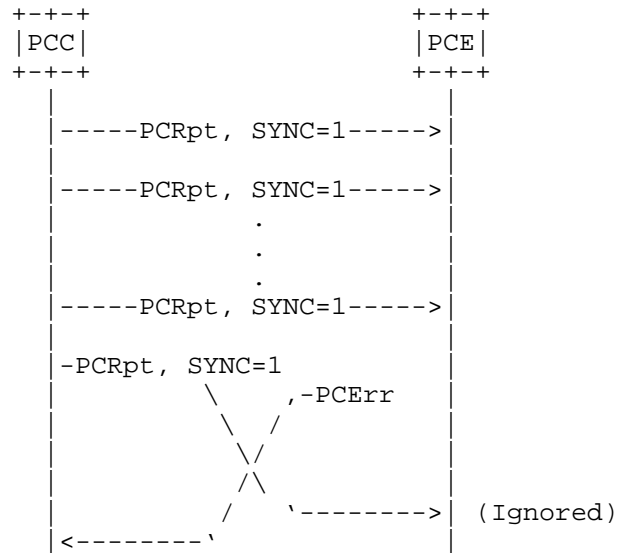


Figure 2: Failed state synchronization (PCE failure)

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

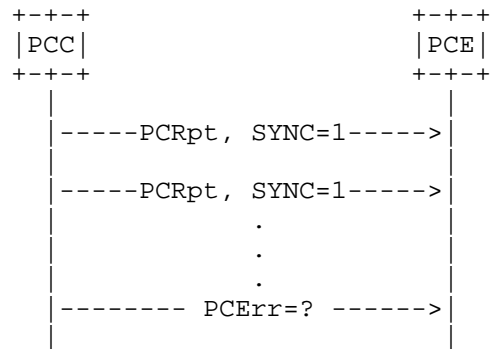


Figure 3: Failed state synchronization (PCC failure)

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.

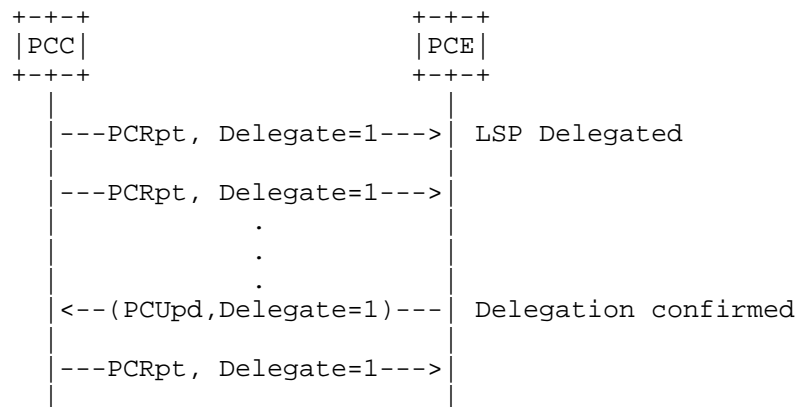


Figure 4: Delegating an LSP

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.

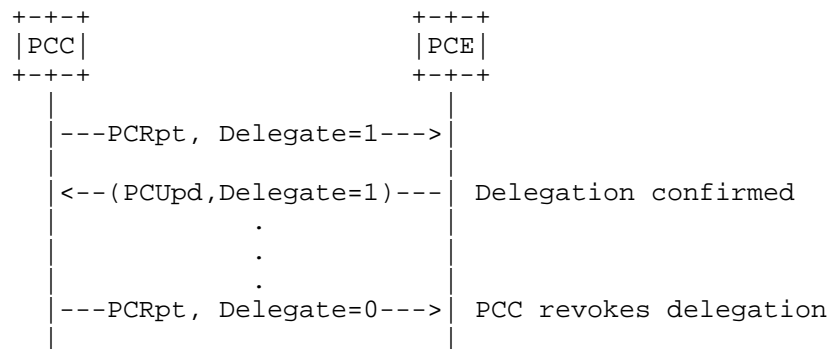


Figure 5: Revoking a Delegation

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).

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 Redelegating 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 Redelegating Timeout, both the delegation state and the LSP state are kept intact.

If the PCE crashes but does not recover within the Redelegating 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 Redelegating 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

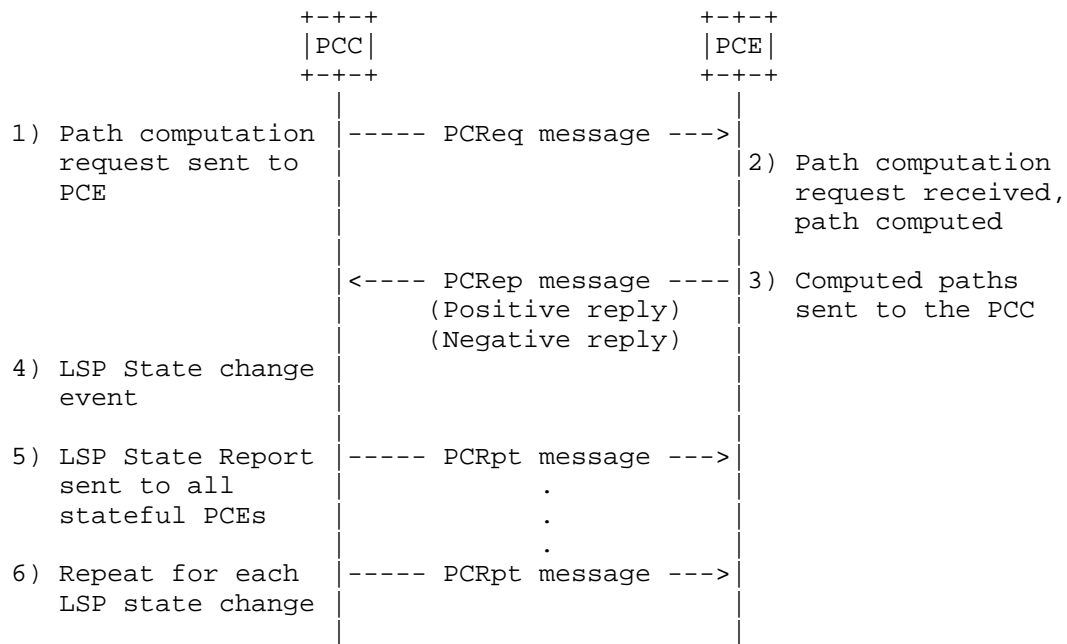


Figure 7: Passive Stateful PCE Path Computation Request/Response

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

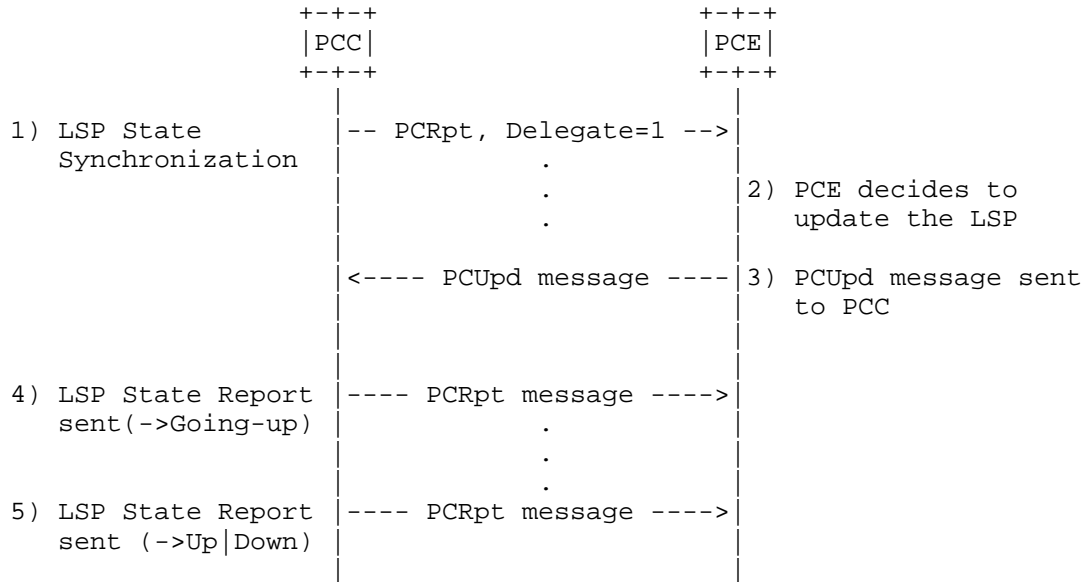


Figure 8: Active Stateful PCE

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:

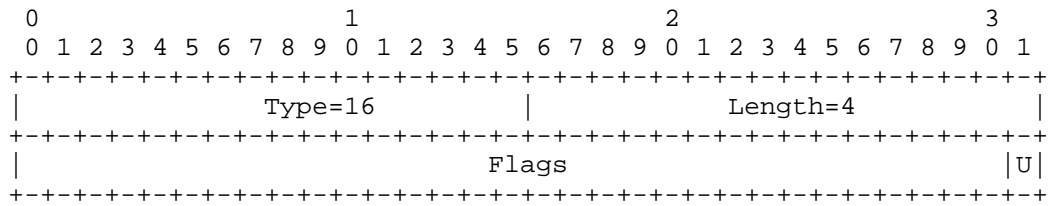


Figure 9: STATEFUL-PCE-CAPABILITY TLV format

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:

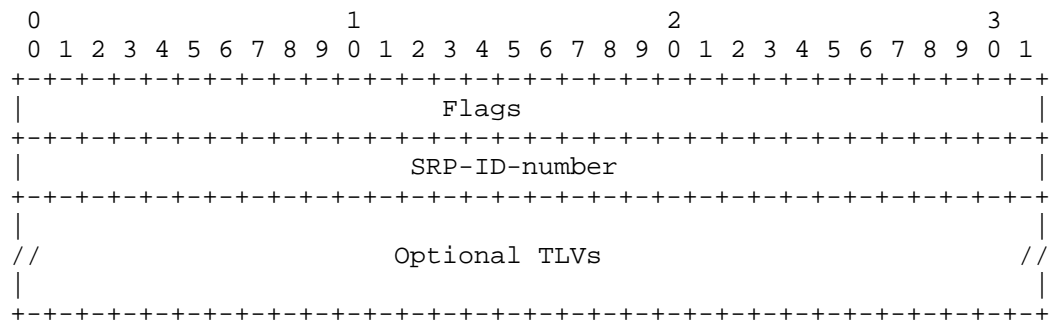


Figure 10: The SRP Object format

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:

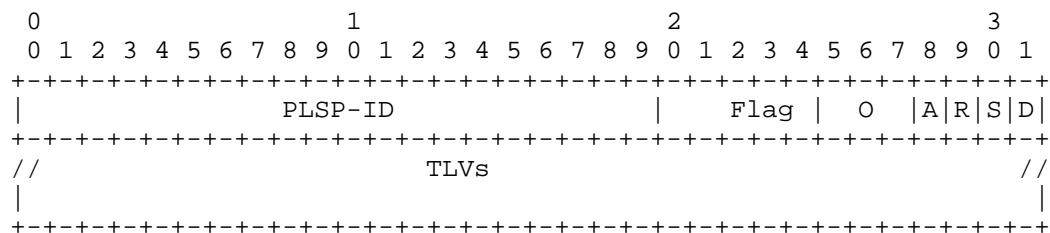


Figure 11: The LSP Object format

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 times. 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:

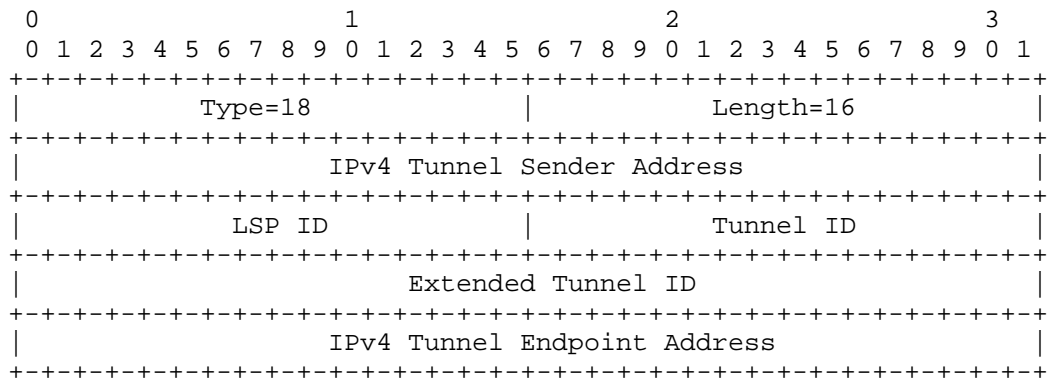


Figure 12: IPV4-LSP-IDENTIFIERS TLV format

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:

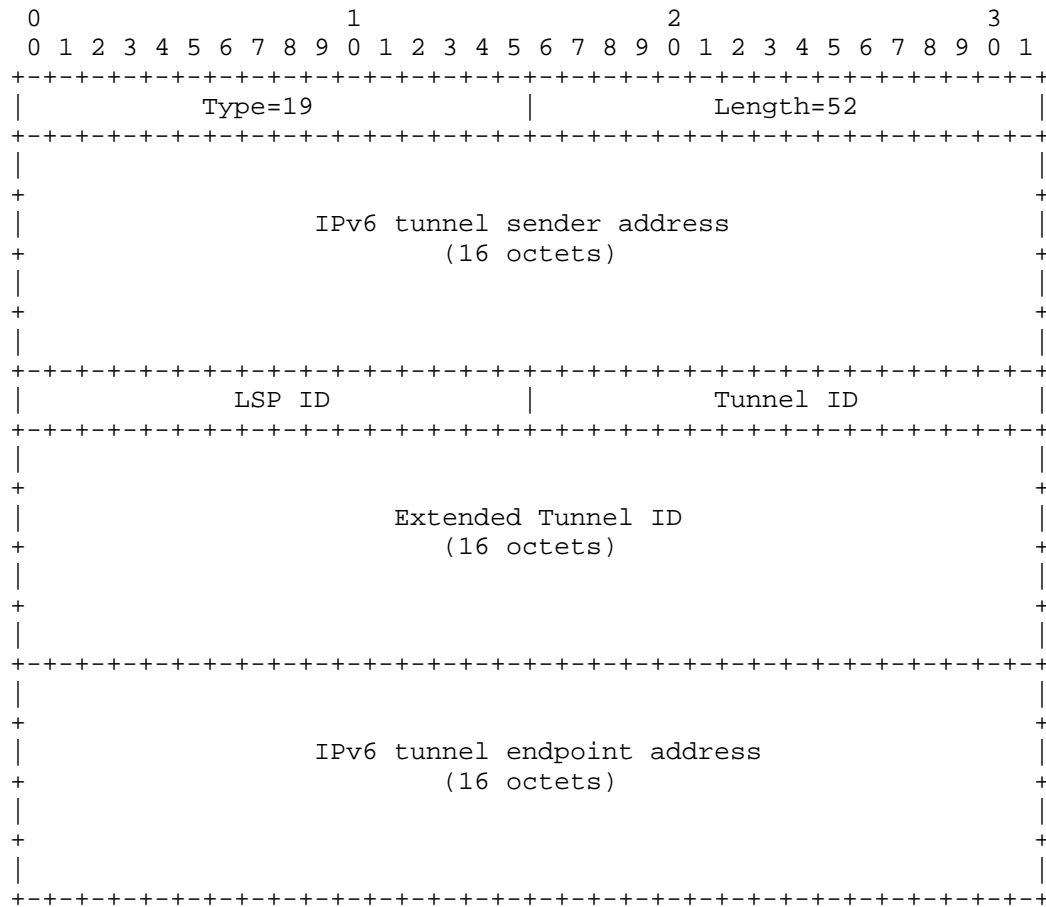


Figure 13: IPV6-LSP-IDENTIFIERS TLV format

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:

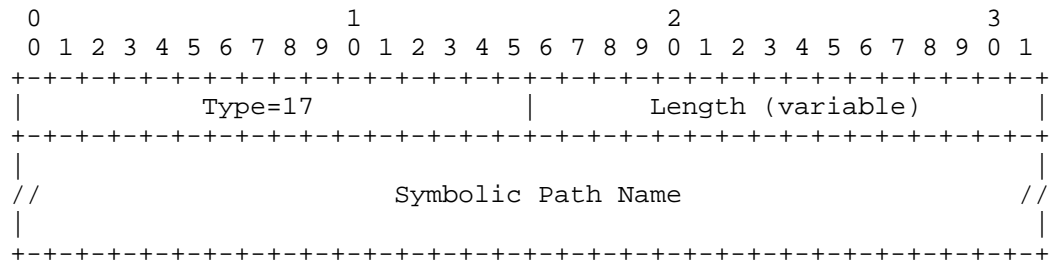


Figure 14: SYMBOLIC-PATH-NAME TLV format

```
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:

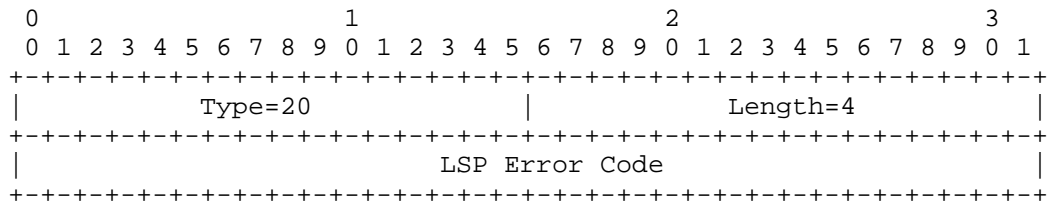


Figure 15: LSP-ERROR-CODE TLV format

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:

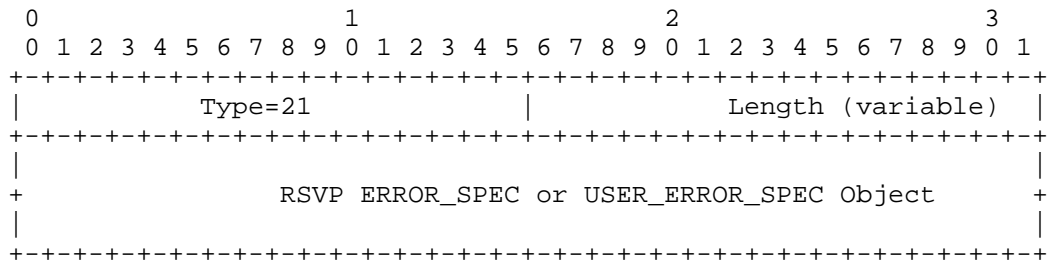


Figure 16: RSVP-ERROR-SPEC TLV format

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

Ina Minei
Google, Inc.
1600 Amphitheatre Parkway
Mountain View, CA 94043
US

Email: inaminei@google.com

Jan Medved
Cisco Systems, Inc.
170 West Tasman Dr.
San Jose, CA 95134
US

Email: jmedved@cisco.com

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

Email: robert.varga@pantheon.tech

PCE Working Group
Internet-Draft
Intended status: Informational
Expires: May 4, 2017

X. Zhang, Ed.
Huawei Technologies
I. Minei, Ed.
Google, Inc.
October 31, 2016

Applicability of a Stateful Path Computation Element (PCE)
draft-ietf-pce-stateful-pce-app-08

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, as well as its challenges and limitations through a number of use cases. PCE Communication 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 May 4, 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	2
2. Terminology	3
3. Application Scenarios	4
3.1. Optimization of LSP Placement	4
3.1.1. Throughput Maximization and Bin Packing	5
3.1.2. Deadlock	7
3.1.3. Minimum Perturbation	8
3.1.4. Predictability	9
3.2. Auto-bandwidth Adjustment	11
3.3. Bandwidth Scheduling	11
3.4. Recovery	12
3.4.1. Protection	12
3.4.2. Restoration	13
3.4.3. SRLG Diversity	14
3.5. Maintenance of Virtual Network Topology (VNT)	15
3.6. LSP Re-optimization	15
3.7. Resource Defragmentation	16
3.8. Point-to-Multi-Point Applications	17
3.9. Impairment-Aware Routing and Wavelength Assignment (IA-RWA)	17
4. Deployment Considerations	18
4.1. Multi-PCE Deployments	18
4.2. LSP State Synchronization	19
4.3. PCE Survivability	19
5. Security Considerations	19
6. IANA Considerations	20
7. Contributing Authors	20
8. Acknowledgements	21
9. References	21
9.1. Normative References	21
9.2. Informative References	22
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) for interaction between a Path Computation Client (PCC) and a PCE, or between two PCEs, enabling computation of TE LSPs.

As per [RFC4655], a PCE can be either stateful or stateless. 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. This state information allows the PCE to compute constrained paths while considering individual LSPs and their inter-dependency. However, this requires reliable state synchronization mechanisms between the PCE and the network, between the PCE and the PCCs, 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 defines the following terms:

Stateful PCE: a PCE that has access to not only the network state, but also to the set of active paths and their reserved resources for its computations. A stateful PCE might also retain information regarding LSPs under construction in order to reduce churn and resource contention. The additional 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.

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

Active Stateful PCE:: a PCE that may issue recommendations to the network. For example, an Active Stateful PCE may utilize the Delegation mechanism to update LSP parameters in those PCCs that delegated control over their LSPs to the PCE.

Delegation: an operation to grant a PCE temporary rights to modify a subset of LSP parameters on one or more PCC's LSPs. LSPs are delegated from a PCC to a PCE, and are referred to as delegated LSPs. The PCC that owns the PCE state for the LSP has the right to delegate it. An LSP is owned by a single PCC at any given point in time. For intra-domain LSPs, this PCC should be the LSP head end.

LSP State Database: information about all LSPs and their attributes.

PCE Initiation: a PCE, assuming LSP delegation granted by default, can issue recommendations to the network.

Minimum Cut Set: the minimum set of links for a specific source destination pair which, when removed from the network, results 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. Application Scenarios

In the following sections, several use cases are described, showcasing scenarios that benefit from the deployment of a stateful PCE.

3.1. Optimization of LSP Placement

The following use cases demonstrate a need for visibility into global LSP states 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 maintenance, and focuses specifically on the active stateful PCE model of operation.

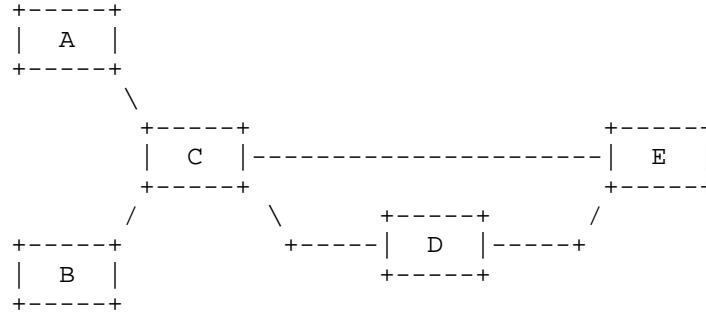


Figure 1: Reference topology 1

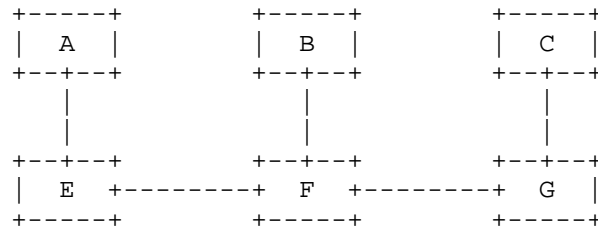


Figure 2: Reference topology 2

3.1.1. Throughput Maximization and Bin Packing

Because LSP attribute changes in [RFC5440] are driven by Path Computation Request (PCReq) messages under control of a PCC's local timers, the sequence of resource 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

3.1.2. 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 label edge router (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 (for example for tunnels that are dynamically resized based on current traffic) and b) there is currently no efficient commonly available northbound interface for dynamic configuration of per LSP ingress admission control.

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, thus 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

3.1.3. 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 computing both routes at the same time. The advantages of using a stateful PCE over exploiting a stateless PCE via Global Concurrent Optimization(GCO) are three folds. First is the ability to accommodate concurrent path computation from different PCCs. Second is the reduction of control plane overhead since the stateful PCE has the route information of the affected LSPs. Thirdly, the stateful PCE can use the LSP-DB to further optimize the placement of LSPs. This will ensure placement of the more important LSP along the shortest path, avoiding the setup and subsequent preemption of the lower priority LSP. Similarly, when a new higher priority LSP which requires preemption of existing lower priority LSP(s), a stateful PCE can determine the minimum number of lower priority LSP(s) to reroute using the make-before-break (MBB) mechanism without disrupting any service and then set up the higher priority LSP.

3.1.4. 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 are routed in both cases, but along very different paths. This would be a challenge if reliable simulation of the network is attempted. An active stateful PCE can solve this through control over LSP ordering. Based on triggers such as a failure or an optimization trigger, the PCE can order the computations and path setup in a deterministic way.

3.2. Auto-bandwidth Adjustment

The bandwidth requirement of LSPs often change over time, requiring resizing the LSP. In most implementations available today, the head-end node performs this function by monitoring the actual bandwidth usage, triggering a recomputation and ressignaling 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. With a stateless PCE, the head-end node needs to provide the current used bandwidth and the route information via path computation request messages. 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.

3.3. 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 network management system (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. With PCE initiation capability, a PCE can trigger the setup and deletion of scheduled requests in a centralized manner, without modification of existing head-end behaviors, by notifying the PCCs to set up or tear down the paths.

3.4. 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.

3.4.1. Protection

If a PCC can specify in a request whether the computation is for a working path or for protection, and a PCC can report the resource as a working or protection path, then the following text applies. A PCC can send multiple requests to the PCE, asking for two LSPs and use them as working and backup paths 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 this information as a constraint in the path computation request. A passive stateful PCE can more easily accommodate this need using the information stored in its LSP-DB. Furthermore, an active stateful PCE can help with (re)-optimization of protection resource sharing as well as LSP maintenance operation with fewer impact on protection resources.

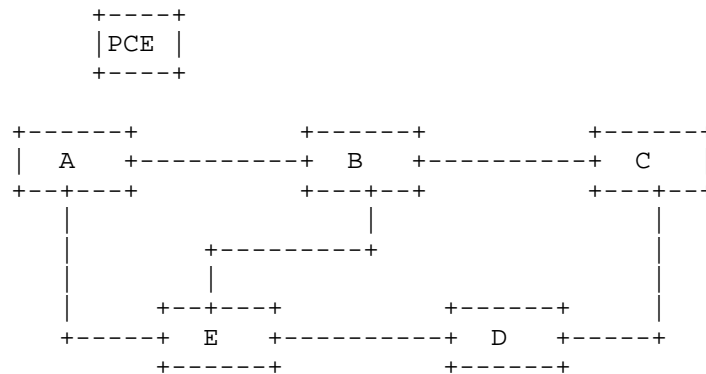


Figure 3: Reference topology 3

For example, in the network depicted in Figure 3, suppose there exists LSP1 with working path LSP1_working following A->E and with backup path LSP1_backup following A->B->E. A request arrives asking for a working and backup path pair to be computed for LSP2 from B to E. If the PCE decides LSP2_working follows B->A->E, then the backup path LSP2_backup should not share the same protection resource with LSP1 since LSP2 shares part of its resource (specifically A->E) with LSP1 (i.e., these two LSPs are in the same shared risk group). There is no such constraint if B->C->D->E is chosen for LSP2 working.

If a stateless PCE is used, the head node B needs to be aware of the existence of LSPs which share the route of LSP2_working and of the details of their protection resources. B must pass this information to the PCE as a constraint so as to request a path with diversity. Alternatively, a stateless PCE may be able to compute Shared Risk Link Group (SRLG)-diversified paths if TED is extended so that it includes the SRLG information that are protected by a given backup resource, but at the expense of a high complexity in routing. On the other hand, a stateful PCE can get the LSPs information by itself given that the LSP identifier(s) 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.

3.4.2. Restoration

In case of a link failure, such as a 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 Exclude Route Object (XRO) with Fail (F) bit set, together with the record route object (RRO) containing the current route information, as specified in [RFC5521].

If a stateless PCE is used, 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 passive stateful PCE is used to fulfill this task, the procedure can be simplified. The PCCs reporting the failures can include LSP identifiers instead of detailed information and the PCE can find relevant LSP information by inspecting the LSP-DB. Moreover, the PCE 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 passive 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 can be achieved, such as better resource usage or minimal probability of blocking upcoming new rerouting requests sent as a result of the link failure.

If the target is to avoid resource contention within the time-window of high number of LSP rerouting 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 alleviated.

3.4.3. 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, the PCC can specify, as constraints to the path computation a set of SRLGs using the Exclude 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 passive 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. Similarly, a passive stateful PCE can also accommodate disjointness using other constraints, such as link, node or path segment etc.

3.5. 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. Hence, when a stateless PCE cannot find the route for a request based on the upper layer topology information, it does not have enough information to decide whether to set up or remove a TE link or not, which then can result in non-optimal usage of resource. On the other hand, a passive stateful PCE can make a better 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]. Furthermore, given the active capability, the stateful PCE can issue VNT modification suggestions in order to accommodate path setup requests or re-optimize resource usage across layers.

3.6. 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 passive stateful PCE, given the LSP state information in the LSP database, the process of dynamic optimization of network resources can be simplified without requiring the PCC to supply detailed LSP state information. Moreover, an active stateful PCE can even make the process automated by triggering the request since a stateful PCE can maintain information for all LSPs that are in the process of being set up and 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 stateful PCE can

also determine which LSP should be re-optimized based on network events. For example, when a LSP is torn down, its resources are freed. This can trigger the stateful PCE to automatically determine which LSP should be reoptimized so that the recently freed resources may be allocated to it.

A special case of LSP re-optimization is 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 network nodes, 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.

3.7. Resource Defragmentation

If LSPs are dynamically allocated and released over time, the resource becomes fragmented. In networks with link bundle, 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.

Another case of particular interest is the optical spectrum defragmentation in flexible grid networks. In Flexible grid networks [RFC7698], LSPs with different optical spectrum sizes (such as 12.5GHz, 25GHz etc.) can co-exist so as to accommodate the services with different bandwidth requests. Therefore, even if the overall spectrum size can meet the service request, it may not be usable if the available spectrum resource is not contiguous, but rather

fragmented into smaller pieces. Thus, with the help of existing LSP state information, a stateful PCE can make the resource grouped together to be usable. Moreover, a stateful PCE can proactively choose routes for upcoming path requests to reduce the chance of spectrum fragmentation.

3.8. Point-to-Multi-Point Applications

PCE has been identified as an appropriate technology for the determination of the paths of point-to-multipoint (P2MP) TE LSPs [RFC5671]. The application scenarios and use-cases described in Section 3.1, Section 3.4 and Section 3.6 are also applicable to P2MP TE LSPs.

In addition to these, the stateful nature of a PCE simplifies the information conveyed in PCEP messages since it is possible to refer to the LSPs via an identifier. For P2MP, this is an added advantage, where the size of the PCEP message is much larger. In case of stateless PCEs, modification of a P2MP tree requires encoding of all leaves along with the paths in PCReq message. But using a stateful PCE with P2MP capability, the PCEP message can be used to convey only the modifications (the other information can be retrieved from the identifier via the LSP-DB).

3.9. Impairment-Aware Routing and Wavelength Assignment (IA-RWA)

In Wavelength Switched Optical Networks (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 10Gb/s signals, mixed 10Gb/s and 40Gb/s signals, or mixed 40Gb/s and 100Gb/s signals.

IA-RWA in WSONs refers to the 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., [RFC7688] and [RFC7580], 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.

4. Deployment Considerations

This section discusses general issues with stateful PCE deployments, and identifies areas where additional protocol extensions and procedures are needed to address them. Definitions of protocol mechanisms are beyond the scope of this document.

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.

Multiple stateful PCEs can co-exist in the same network. These PCEs may provide redundancy for load sharing, resilience, or partitioning of computation features. Regardless of the reason for multiple PCEs, an LSP is only delegated to one of the PCEs at any given point in time. However, an LSP can be re-delegated between PCEs, for example when a PCE fails. [RFC7399] discusses various approaches for synchronizing state among the PCEs when multiple PCEs are used for load sharing or backup and compute LSPs for the same network.

4.2. LSP State Synchronization

The LSP-DB is populated using information received from the PCC. Because the accuracy of the computations depends on the accuracy of the databases used, it is worth noting that the PCE view lags behind the true state of the network, because the updates must reach the PCE from the network. Thus, the use of stateful PCE reduces but cannot eliminate the possibility of crankbacks, nor can it guarantee optimal computations all the time. [RFC7399] discusses these limitations and potential ways to alleviate them.

In case of multiple PCEs with different capabilities, co-existing in the same network, such as a passive stateful PCE and an active stateful PCE, it is useful to refer to a LSP, be it delegated or not, by a unique identifier instead of providing detailed information (e.g., route, bandwidth etc.) associated with it, when these PCEs cooperate on path computation, such as for load sharing.

4.3. PCE Survivability

For a stateful PCE, an important issue is to get the LSP state information resynchronized after a restart. LSP state synchronization procedures can be applied equally to a network node or another PCE, allowing multiple ways of re-acquiring the LSP database on a restart. 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. A hybrid approach where the bulk of the state is recovered locally, and a small amount of state is reacquired from the network, is also possible. Note that locally recovering the state would still require some degree of resynchronization to ensure that the recovered state is indeed up-to-date. Depending on the resynchronization mechanism used, there may be an additional load on the PCE, and there may be a delay in reaching the synchronized state, which may negatively affect survivability. Different resynchronization methods are suited for different deployments and objectives.

5. Security Considerations

This document describes general considerations for a stateful PCE deployment and examines its applicability and benefits, as well as its challenges and limitations through a number of use cases. No new protocol extensions to PCEP are defined in this document.

The PCEP extensions in support of the stateful PCE and the delegation of path control ability can result in more information and control being available for a hypothetical adversary and a number of

additional attack surfaces which must be protected. This includes but not limited to the authentication and encryption of PCEP sessions, snooping of the state of the LSPs active in the network etc. Therefore, documents where the PCEP protocol extensions are defined need to consider the issues and risks associated with a stateful PCE.

6. IANA Considerations

This document does not require any IANA action.

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
Email: edward.crabbe@gmail.com

Dhruv Dhody
Huawei Technology
Leela Palace
Bangalore, Karnataka 560008
INDIA
Email: dhruv.dhody@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: leeyoung@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, JP Vasseur and Ravi Torvi for the useful comments and discussions.

9. References

9.1. Normative 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>>.
- [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>>.

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

9.2. 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, DOI 10.17487/RFC3209, December 2001, <<http://www.rfc-editor.org/info/rfc3209>>.
- [RFC4427] Mannie, E., Ed. and D. Papadimitriou, Ed., "Recovery (Protection and Restoration) Terminology for Generalized Multi-Protocol Label Switching (GMPLS)", RFC 4427, DOI 10.17487/RFC4427, March 2006, <<http://www.rfc-editor.org/info/rfc4427>>.
- [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>>.
- [RFC5212] Shiomoto, 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, DOI 10.17487/RFC5212, July 2008, <<http://www.rfc-editor.org/info/rfc5212>>.
- [RFC5521] Oki, E., Takeda, T., and A. Farrel, "Extensions to the Path Computation Element Communication Protocol (PCEP) for Route Exclusions", RFC 5521, DOI 10.17487/RFC5521, April 2009, <<http://www.rfc-editor.org/info/rfc5521>>.
- [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, DOI 10.17487/RFC5557, July 2009, <<http://www.rfc-editor.org/info/rfc5557>>.
- [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>>.

- [RFC5671] Yasukawa, S. and A. Farrel, Ed., "Applicability of the Path Computation Element (PCE) to Point-to-Multipoint (P2MP) MPLS and GMPLS Traffic Engineering (TE)", RFC 5671, DOI 10.17487/RFC5671, October 2009, <<http://www.rfc-editor.org/info/rfc5671>>.
- [RFC6163] Lee, Y., Ed., Bernstein, G., Ed., and W. Imajuku, "Framework for GMPLS and Path Computation Element (PCE) Control of Wavelength Switched Optical Networks (WSOs)", RFC 6163, DOI 10.17487/RFC6163, April 2011, <<http://www.rfc-editor.org/info/rfc6163>>.
- [RFC7580] Zhang, F., Lee, Y., Han, J., Bernstein, G., and Y. Xu, "OSPF-TE Extensions for General Network Element Constraints", RFC 7580, DOI 10.17487/RFC7580, June 2015, <<http://www.rfc-editor.org/info/rfc7580>>.
- [RFC7688] Lee, Y., Ed. and G. Bernstein, Ed., "GMPLS OSPF Enhancement for Signal and Network Element Compatibility for Wavelength Switched Optical Networks", RFC 7688, DOI 10.17487/RFC7688, November 2015, <<http://www.rfc-editor.org/info/rfc7688>>.
- [RFC7698] Gonzalez de Dios, O., Ed., Casellas, R., Ed., Zhang, F., Fu, X., Ceccarelli, D., and I. Hussain, "Framework and Requirements for GMPLS-Based Control of Flexi-Grid Dense Wavelength Division Multiplexing (DWDM) Networks", RFC 7698, DOI 10.17487/RFC7698, November 2015, <<http://www.rfc-editor.org/info/rfc7698>>.

Authors' Addresses

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

Email: zhang.xian@huawei.com

Ina Minei (editor)
Google, Inc.
1600 Amphitheatre Parkway
Mountain View, CA 94043
US

Email: inaminei@google.com

PCE Working Group
Internet Draft
Intended status: Informational

Young Lee
Xian Zhang
Haomian Zheng
Huawei
Guoying Zhang
CATR

Expires: April 21 2014

October 18, 2013

Application-oriented Stateful PCE Architecture and Use-cases for
Transport Networks

draft-lee-pce-app-oriented-arch-00.txt

Status of this Memo

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

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

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

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

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

This Internet-Draft will expire on April 21, 2013.

Abstract

This draft presents an application-oriented stateful PCE architecture for transport networks. Under this architecture, several use cases are described.

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.....	2
2. Terminology.....	3
3. Architecture and Key Features.....	3
4. Use-cases.....	5
4.1. Dynamic Data Center Network Interconnection.....	6
4.2. Packet-Optical Integration (POI).....	6
4.3. Virtual Network Service (VNS).....	6
4.4. Time-based Scheduling.....	7
5. References.....	7
5.1. Informative References.....	7
6. Authors' Addresses	8
7. Acknowledgment.....	9

1. Introduction

With the emerging applications requiring large bandwidth and dynamic provisioning, such as Data Center Interconnection(DCI), cloud bursting and so on, the traditional transport network architecture is limited as it only provides "dumb pipe" services. These services lack the flexibility for operation and management. In order to support the demands, including large bandwidth, low service latency as well as dynamic and flexible resource allocation, transport networks may need to be enhanced architecturally such that it could be aware of application requirements in a dynamic fashion. The Path Computation Elements (PCE) architecture and the corresponding protocol extensions provide a mechanism that enables path computation for transport network. As specified in [RFC4655], a PCE supports the request for path computation issued by a Path

Computation Client (PCC). When the PCC is external to the PCE, a communication protocol, i.e., PCE Protocol (PCEP), is required to support the path computation request/reply process. Furthermore, extensions to PCEP are proposed in [PCE-S] , [PCE-I], and [PCE-S-GMPLS] to enable stateful control over networks including transport networks.

This draft provides an application-oriented stateful PCE architecture for transport networks. In particular, this architecture introduces transport network controller (TNC) component in which transport PCE plays a central role. Given the high demands from applications, an interface between the transport network controller and the application client controller is also introduced to enable the communication function between these entities. The application client controller is a special type of PCC with respect to PCE capability within the transport network controller. This interface and its communication mechanism between the application client controller and the transport network controller enables operation of the transport network with more flexibility. Specifically, in a larger-scale transport network with multiple layers or multiple domains, the communication mechanism between different PCEs and the application client controllers is very important to satisfy the request from the application stratum. Current PCEP can provide communication between PCE and PCCs, and further extensions to PCEP may be desirable to cooperate with new types of PCCs such as application client controllers.

2. Terminology

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

3. Architecture and Key Features

In this draft, a PCE-centric architecture which supports application-oriented transport network is defined. The architecture is illustrated in Figure 1. The functions of each architectural component are described. And then interfaces between the stateful PCE and the other functional blocks in the transport network are defined.

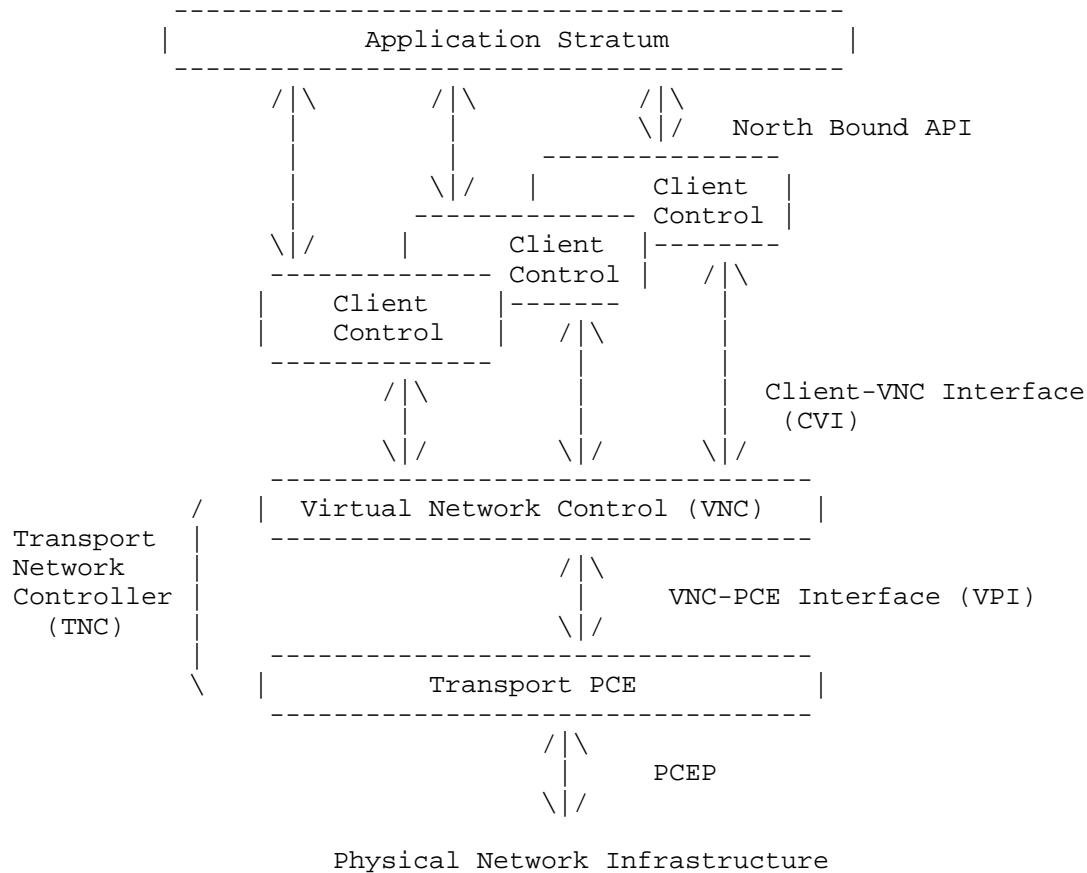


Figure 1: Application-oriented PCE Architecture for Transport Network

Transport Network Controller (TNC) in Figure 1 is the core of the application-oriented PCE architecture for transport network. It is built around the Transport PCE and provides additional functions that facilitate multi-layer control, virtual network service control and other functionalities such as topology abstraction via the Virtual Network Control (VNC) block. The VNC interfaces can be different types of client controllers, such as packet network

controllers, data center provider controllers, enterprise network controllers, virtual service provider controllers, etc. The VNC provides network control function virtualization to the PCE and to the clients via the VNC-PCE Interface (VPI) and the Client-VNC Interface (CVI), respectively. The VNC allows the clients (via their client controllers) to program their client-defined virtual network services (VNS) over the CVI. The VNC also provides abstract network topology for each client based on the network resources allocated to the client. In order to facilitate this capability, the VNC needs to communicate with the PCE via the VPI interface. The VPI can be an internal interface from an implementation standpoint. In this draft, it is assumed an external entity from the PCE. The VNC is a PCC to the PCE.

The VNC provides control plane function virtualization over programmable interfaces such as virtual network path computation and optimization, topology abstraction hiding details of physical topology while supporting service-specific objectives the clients demand, maintaining virtual network service instances and the states, policy enforcement for virtual network services. See [NCFV] for details of control function virtualization concept. With this evolutionary architecture built on top of transport PCE, a number of challenging use-cases can be supported. Transport PCE is a stateful PCE and supports all the generic stateful PCE functions as described in [PCE-S] and [PCE-S-GMPLS].

The CVI is an external interface with respect to the transport network controller (TNC). Client controller is an external client. Figure 1 shows that there are multiple client controls which are independent to each other and that each client supports various business applications over its NB API. There are layered client-server relationships in this architecture. As various applications are clients to client control, client control itself is also a client to virtual network control; likewise, virtual network control is also a client to physical network control. This layered relationship is important in protocol definition work on NB API, CVI and VPI interfaces as this allows third-party software developers to program client control and virtual network control functions in such a way to create, modify and delete virtual network services.

4. Use-cases

This section provides a number of use-cases to which the architecture discussed in Section 3 is applied.

4.1. Dynamic Data Center Network Interconnection

In the context of multiple data center networks where there is a need to move large data dynamically from one location to other location(s), data center network controller is a type of client controller that coordinates with the virtual network controller (VNC). This coordination across data center client controller and the VNC allows multiple instances of inter data center connections need for different applications.

For each application, the VNC keeps the instance and creates an abstracted network topology based on the network resources allocated to a particular application. The data center client controller has the view of this abstracted network topology and is given a full control of how to use the allocated virtual resources.

The topology abstraction created by the VNC for the client is based on the transport PCE's real network resource information and is needed to be filtered via the VNC's filtering mechanism based on contract, policy and security.

The VNC interlays client control's request for inter data center connection and converts into a PC request to the PCE. Then a PCE instantiates a network path via its provisioning mechanism described in [PCE-I].

4.2. Packet-Optical Integration (POI)

Client controller can also be a router network controller that needs transport network interconnections. The router network controller can request different connection services from the transport network based on different QoS needs.

Note that this POI use-case is different from multi-layer PCE work [RFC5623] in that it allows more flexible interactions and more granular level of abstracted network topologies than tunnel-based virtual network topology.

4.3. Virtual Network Service (VNS)

Virtual Network Service is instantiated by the client control via the CVI. As client control directly interfaces the application stratum, it understands multiple application requirements and their service needs. It is assumed that client control and VNC have the common knowledge on the end-point interfaces based on their business negotiation prior to service instantiation. End-point interfaces refer to client-network physical interfaces that connect client

premise equipment to network provider equipment. The different level of topology abstractions can be provided by the VNC topology abstraction engine based on physical topology base maintained by the PNC.

The level of topology abstraction is expressed in terms of the number of virtual network elements (VNEs) and virtual links (VLs). As different client has different control/application needs, abstracted topologies for a certain client can show much less degree of abstraction. The level of topology abstraction is determined by the policy (e.g., the granularity level) placed for the client and/or the path computation results by the PNC's PCE. The finer granularity the abstraction topology is, the more control is given to the client control. For instance, if the client is a third-party virtual service broker/provider, then it would desire much more sophisticated control of virtual network resources to support differing application needs. On the other hand, if the client were only to support simple tunnel services to its application, then abstracted topology for such client is a simple abstracted topology with a set of end-point tunnels.

4.4. Time-based Scheduling

Transport services with time constraints are another highly-demanded task in the network. In this scenario, a client controller can request to reserve some bandwidth for future use. This 'time-based' service needs to be considered together with the traffic Engineering Database (TED) and Label Switched Path Database (LSPD). PCE will compute the scheduled network resource for this 'time-based' service, and reserve such resources for future use.

In this scenario, the LSPD contains two categories of LSP information, current LSP in use and scheduled LSP. These two groups of LSP can be included in a single LSPD or two separate ones, with internal interface to PCE. PCEP should also be extended to include the scheduled information for service requests, such as proposed in [Time-based]. With these extensions, the PCC (for example, application stratum) can generate the path computation request.

5. References

5.1. Informative References

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

- [RFC5440] Vasseur, J. P. and J. L. Le Roux, "Path Computation Element (PCE) Communication Protocol (PCEP)", RFC 5440, March 2009.
- [RFC5623] Oki, E., Takeda, T., et al, "Framework for PCE-Based Inter-Layer MPLS and GMPLS Traffic Engineering", RFC 5623, September 2009.
- [PCE-S] Crabbe, E., Medved, J., Minei, I., and R. Varga, "PCEP Extensions for Stateful PCE", draft-ietf-pce-stateful-pce, work in progress. [PCE-I] Crabbe, E., Minei, I., Sivabalan, S., and Varga, R., "PCEP Extensions for PCE-initiated LSP Setup in a Stateful PCE Model", draft-crabbe-pce-pce-initiated-lsp, work in progress.
- [PCE-S-GMPLS] Zhang, X., Lee, Y., Casellas, R., Gonzalez de Dios, O., "Path Computation Element (PCE) Protocol Extensions for Stateful PCE Usage in GMPLS-controlled Networks", draft-zhang-pce-pcep-stateful-pce-gmpls-03.txt, work in progress.
- [NCVF] Lee, Y. Bernstein, G., So, N., Fang L., and Ceccarelli, D. "Network Control Function Virtualization for Transport SDN", draft-lee-network-control-function-virtualization, work in progress.
- [Time-based] Zhang, X., Lee, Y., Casellas, R., Ganzalez, O., "Stateful Path Computation Element (PCE) for Time-based Scheduling", draft-zhang-pce-stateful-time-based-scheduling-00, work in progress.

6. Authors' Addresses

Young Lee
Huawei Technologies
5340 Legacy Dr.
Plano, TX 75023, USA
Phone: (469)277-5838
Email: leeyoung@huawei.com

Xian Zhang

Huawei Technologies
Email: zhang.xian@huawei.com

Haomian Zheng
Huawei Technologies
Email: Zhenghaomian@huawei.com

Guoying Zhang
China Academy of Telecommunication Research of MII
11 Yue Tan Nan Jie Beijing, P.R.China
Phone: +86-10-68094272
Email: zhangguoying@mail.ritt.com.cn

7. Acknowledgment

Intellectual Property

The IETF Trust takes no position regarding the validity or scope of any Intellectual Property Rights or other rights that might be claimed to pertain to the implementation or use of the technology described in any IETF Document or the extent to which any license under such rights might or might not be available; nor does it represent that it has made any independent effort to identify any such rights.

Copies of Intellectual Property disclosures made to the IETF Secretariat and any assurances of licenses to be made available, or the result of an attempt made to obtain a general license or permission for the use of such proprietary rights by implementers or users of this specification can be obtained from the IETF on-line IPR repository at <http://www.ietf.org/ipr>

The IETF invites any interested party to bring to its attention any copyrights, patents or patent applications, or other proprietary rights that may cover technology that may be required to implement any standard or specification contained in an IETF Document. Please address the information to the IETF at ietf-ipr@ietf.org.

The definitive version of an IETF Document is that published by, or under the auspices of, the IETF. Versions of IETF Documents that are published by third parties, including those that are translated into

other languages, should not be considered to be definitive versions of IETF Documents. The definitive version of these Legal Provisions is that published by, or under the auspices of, the IETF. Versions of these Legal Provisions that are published by third parties, including those that are translated into other languages, should not be considered to be definitive versions of these Legal Provisions.

For the avoidance of doubt, each Contributor to the IETF Standards Process licenses each Contribution that he or she makes as part of the IETF Standards Process to the IETF Trust pursuant to the provisions of RFC 5378. No language to the contrary, or terms, conditions or rights that differ from or are inconsistent with the rights and licenses granted under RFC 5378, shall have any effect and shall be null and void, whether published or posted by such Contributor, or included with or in such Contribution.

Disclaimer of Validity

All IETF Documents and the information contained therein are provided on an "AS IS" basis and THE CONTRIBUTOR, THE ORGANIZATION HE/SHE REPRESENTS OR IS SPONSORED BY (IF ANY), THE INTERNET SOCIETY, THE IETF TRUST AND THE INTERNET ENGINEERING TASK FORCE DISCLAIM ALL WARRANTIES, EXPRESS OR IMPLIED, INCLUDING BUT NOT LIMITED TO ANY WARRANTY THAT THE USE OF THE INFORMATION THEREIN WILL NOT INFRINGE ANY RIGHTS OR ANY IMPLIED WARRANTIES OF MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE.

Copyright Notice

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

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

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

V. Lopez
O. Gonzalez de Dios
Telefonica I+D
D. King
Old Dog Consulting
S. Previdi
Cisco Systems, Inc.
October 21, 2013

Traffic Engineering Database dissemination for Hierarchical PCE
scenarios
draft-lopez-pce-hpce-ted-00

Abstract

The PCE architecture is well-defined and may be used to compute the optimal path for LSPs across domains in MPLS-TE and GMPLS networks. The Hierarchical Path Computation Element (H-PCE) [RFC6805] was developed to provide an optimal path when the sequence of domains is not known in advance. The procedure and mechanism for populating the Traffic Engineering Database (TED) with domain topology and link information used in H-PCE-based path computations is open to interpretation. This informational document describes how topology dissemination mechanisms may be used to provide TE information between Parent and Child PCEs (within the H-PCE context). In particular, it describes how BGP-LS might be used to provide inter-domain connectivity. This document is not intended to define new extensions, it demonstrates how existing procedures and mechanisms may be used.

Status of This Memo

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

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

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

This Internet-Draft will expire on April 24, 2014.

Copyright Notice

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

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

Table of Contents

1. Introduction	2
1.1. Parent PCE Domain Topology	3
1.2. Parent PCE TED requirements	3
2. H-PCE Domain Topology Dissemination and Construction Methods	4
3. H-PCE architecture using BGP-LS	5
4. Including Inter-domain connectivity in BGP-LS	8
4.1. Mapping from OSPF-TE	8
4.2. Mapping from ISIS-TE	8
5. BGP considerations	8
6. Manageability Considerations	8
7. Security Considerations	8
8. Acknowledgements	9
9. References	9
9.1. Normative References	9
9.2. Informative References	9
Authors' Addresses	10

1. Introduction

In scenarios with multiple domains in both MPLS-TE and GMPLS networks, the hierarchical Path Computation Element (H-PCE) Architecture, defined in [RFC6805], allows to obtain the optimum end-to-end path. The architecture exploits a hierarchical relation among domains.

[RFC6805] defines the architecture and requirements for the end-to-end path computation across domains. The solution draft for the H-PCE [I-D.draft-ietf-pce-hierarchy-extensions-00] is focused on the PCEP protocol extensions to support such H-PCE procedures, including negotiation of capabilities and errors. However, neither the architecture nor the solution draft specify which mechanism must to

be used to build and populate the parent PCE (pPCE) Traffic Engineering Database (TED).

The H-PCE architecture documents define the minimum content needed in the traffic engineering database required to compute paths. The information required by parent TEDB are identified in [RFC6805] and further elaborated in [I-D.draft-ietf-pce-inter-area-as-applicability-03]. For instance, [RFC6805] and [I-D.draft-ietf-pce-inter-area-as-applicability-03] suggest that BGP-LS could be used as a "northbound" TE advertisement. This means that a PCE does not need to listen IGP in its domain, but its TED is populated by messages received (for example) from a Route Reflector.

This document highlights the applicability of BGP-LS to the dissemination of domain topology within the H-PCE architecture. In particular, it describes how can BGP-LS be used to send the inter-domain connectivity. It also shows how can OSPF-TE and ISIS-TE updates be mapped into BGP-LS.

Note that this document is not intended to define new protocol extensions, it is an informational document and where required it highlights where existing mechanisms and protocols may be applied.

1.1. Parent PCE Domain Topology

The pPCE maintains a domain topology map of the child domains and their interconnectivity. This map does not include any visibility into the child domains. Where inter-domain connectivity is provided by TE links, the capabilities of those links may also be known to the pPCE. The pPCE maintains a TED for the parent domain, the nodes in the parent domain are abstractions of the cPCE domains (connected by real or virtual TE links), but the pPCE domain may also include real nodes and links.

The procedure and protocol mechanism for disseminating and construction of the pPCE TED may be provided using a number of mechanisms, including manually configuring the necessary information or automated using a separate instance of a routing protocol to advertise the domain interconnectivity. Since inter-domain TE links can be advertised by the IGPs operating in the child domains, this information could then be exported to the parent PCE either by the child PCEs or using north-bound export mechanisms.

1.2. Parent PCE TED requirements

The information that would be exchanged includes:

- o Identifier of advertising child PCE.
- o Identifier of PCE's domain.
- o Identifier of the link.
- o TE properties of the link (metrics, bandwidth).
- o Other properties of the link (technology-specific).
- o Identifier of link endpoints.
- o Identifier of adjacent domain.

2. H-PCE Domain Topology Dissemination and Construction Methods

A variety of methods exist to provide are different alternatives so the parent PCE can get the topological information from the child PCEs (cPCEs):

- o Statically configure all inter-domain link and topology information.
- o Membership of an IGP instance. The necessary topological information could be disseminated by joining the IGP instance of each child PCE domain. However, by doing so, it would break the domain confidentiality principles and is subject to scalability issues.
- o PCEP Notification Messages. Another solution is to send the interconnection information between domains using PCEP Notifications (see section 4.8.4 of [RFC6805]). One approach, followed in research work, is embedding in PCEP Notifications the Inter-AS OSPF-TE Link State Advertisements (LSA) to send the Inter-Domain Link information from child PCEs to the parent PCE and to send reachability information (list of end-points in each domain). However, it is argued that the utilization of PCEP to disseminate topology is beyond scope of the protocol.
- o Separate IGP instance. [RFC6805] points out that in models such as ASON it is possible to consider a separate instance of an IGP running within the parent domain where the participating protocol speakers are the nodes directly present in that domain and the PCEs (parent and child PCEs).
- o Use north-bound distribution of TE information. The North-Bound Distribution of Link-State and TE Information using BGP has been recently propose in the IEFT

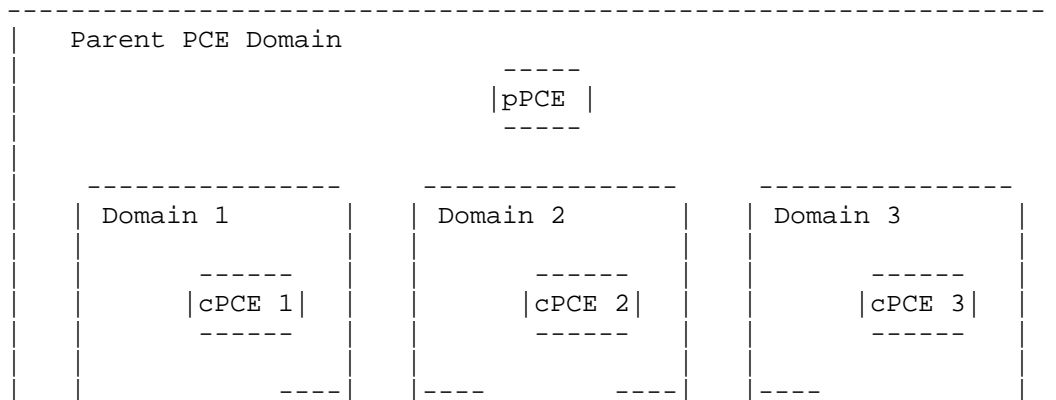
[I-D.draft-ietf-idr-ls-distribution-03]. This approach is known as BGP-LS and defines a mechanism by which links state and traffic engineering information can be collected from networks and exported to external elements using the BGP routing protocol. By using BGP-LS as northbound distribution mechanism, there would be a BGP speaker in each domains that sends the necessary information to a BGP speaker in the parent domain. This architecture is further elaborated in this document.

3. H-PCE architecture using BGP-LS

As mentioned in [I-D.draft-dugeon-pce-ted-reqs-01] PCE has to retrieve Traffic Engineering (TE) information to carry out its path computation. This is required not only for intra-domain information, which can be got using IGP (like OSPF-TE or ISIS-TE), but also for inter-domain information in the Hierarchical PCE (H-PCE) architecture.

Figure 1 shows an example of a H-PCE architecture. In this example, there is a parent PCE and three child PCEs, and they are organized in multiple domains. The parent PCE does not have information of the whole network, but is only aware of the connectivity among the domains and provides coordination to the child PCEs. Figure 2 shows which is the visibility that parent PCE has from the network according to the definition in [RFC6805].

Thanks to this topological information, when there is a request to a child PCE with the destination in another domain, this path request is sent to the parent PCE, which selects a set of candidate domain paths and sends requests to the child PCEs responsible for these domains. Then, the parent PCE selects the best solution and it is transmitted to the source PCE.



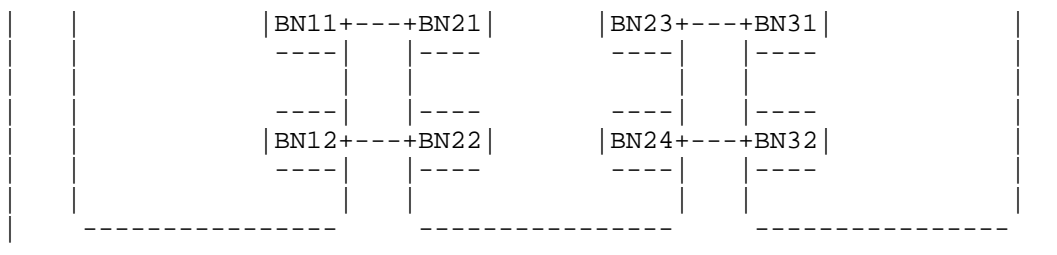


Figure 1: Example of Hierarchical PCE architecture

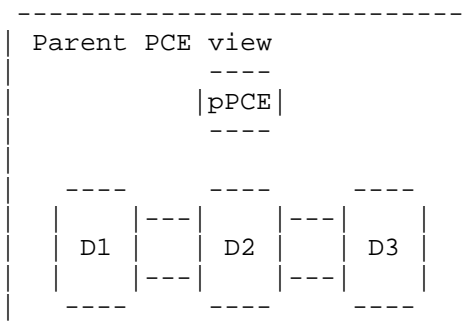
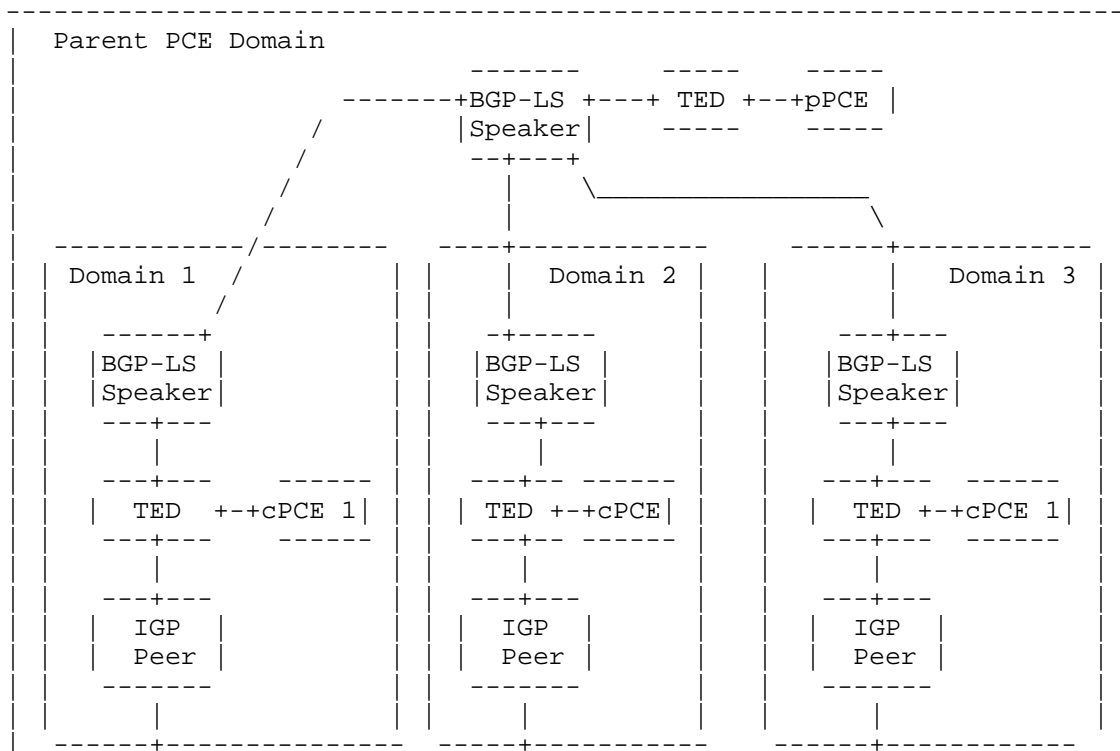


Figure 2: Parent PCE topology information

Thanks to the dissemination of inter-domain adjacency information from each cPCE to the pPCE, the pPCE can have a view of reachability between the domains. The H-PCE architecture with BGP-LS is shown in Figure 3. Each domain has a cPCE that is able to compute paths in the domain. This child PCE has access to a domain TED, which is built using IGP information. In each domain, a BGP speaker has access to such domain TED and acts as BGP-LS Route Reflector to provide network topology to the pPCE. Next to the pPCE, there is a BGP speaker that maintains a BGP session with each of the BGP speakers in the domains to receive the topology and build the parent TED. A policy can be applied to the BGP-LS speakers to decide which information is sent to its peer speaker. The minimum amount of information that needs to be exchanged is the inter-domain connectivity, including the details of the Traffic Engineering Inter-domain Links [RFC6805]. With this information, the parent PCE is able to have access to a domain topology map and its connectivity. Additionally, the BGP-LS speaker can be configured to send the

complete list of TE Links, including its details. In this case, the parent PCE has access to an extended database, with visibility of both intra-domain and inter-domain information and can compute the sequence of domains with better accuracy. Even, the pPCE could have enough information to compute the whole end-to-end path by itself.

BGP-LS [I-D.draft-ietf-idr-ls-distribution-03] extends the BGP Update messages to advertise link-state topology thanks to new BGP Network Layer Reachability Information (NLRI). The Link State information is sent in two BGP attributes, the MP_REACH (defined in [RFC4670]) and a LINK_STATE attribute (defined in [I-D.draft-ietf-idr-ls-distribution-03]). To describe the inter domain links, in the MP_REACH attribute, a Link NLRI can be used with the local node descriptors the address of the source, and in the remote descriptors, the address of the destination of the link. The Link Descriptors field has a TLV (Link Local/Remote Identifiers), which carries the prefix of the Unnumbered or Numbered Interface. In case of the message informs about an intra-domain link, the standard traffic engineering information is included in the LINK_STATE attribute.



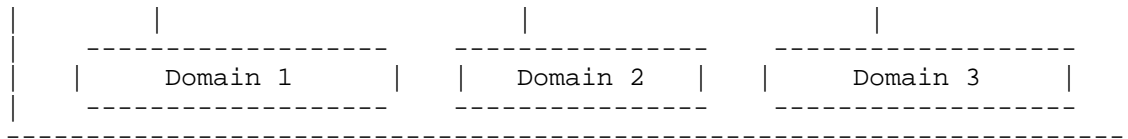


Figure 3: Example of Hierarchical PCE architecture with BGP-LS

4. Including Inter-domain connectivity in BGP-LS

TBD

4.1. Mapping from OSPF-TE

TBD

4.2. Mapping from ISIS-TE

TBD

5. BGP considerations

TBD

- o Supporting BGP-4
- o BGP Speakers
- o Graceful Restart
- o SRLGs
- o Multiprotocol extensions

6. Manageability Considerations

TBD

7. Security Considerations

TBD

8. Acknowledgements

Authors would like to thank Stefano Previdi for his comments.

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.
- [RFC4670] Nelson, D., "RADIUS Accounting Client MIB for IPv6", RFC 4670, August 2006.
- [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.

9.2. Informative References

- [I-D.draft-dugeon-pce-ted-reqs-01]
Dugeon, O., Meuric, J., Douville, R., Casellas, R., and O. Gonzalez de Dios, "Path Computation Element (PCE) Traffic Engineering Database (TED) Requirements", March 2012.
- [I-D.draft-ietf-idr-ls-distribution-03]
Gredler, H., Medved, J., Previdi, S., Farrel, A., and S. Ray, "North-Bound Distribution of Link-State and TE Information using BGP", May 2013.
- [I-D.draft-ietf-pce-hierarchy-extensions-00]
Zhang, F., Zhao, Q., Gonzalez de Dios, O., Casellas, R., and D. King, "Extensions to Path Computation Element Communication Protocol (PCEP) for Hierarchical Path Computation Elements (PCE)", August 2013.
- [I-D.draft-ietf-pce-inter-area-as-applicability-03]
King, D., Meuric, J., Dugeon, O., Zhao, Q., and O. Gonzalez de Dios, "Applicability of the Path Computation Element to Inter-Area and Inter-AS MPLS and GMPLS Traffic Engineering", March 2012.

Authors' Addresses

Victor Lopez
Telefonica I+D
Don Ramon de la Cruz 82-84
Madrid 28045
Spain

Phone: +34913128872
Email: vlopez@tid.es

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

Phone: +34913128832
Email: ogondio@tid.es

Daniel King
Old Dog Consulting
UK

Email: daniel@olddog.co.uk

Stefano Previdi
Cisco Systems, Inc.
Via Del Serafico 200
Rome 00144
IT

Email: sprevidi@cisco.com

Path Computation Element
Internet-Draft
Intended status: Experimental
Expires: April 23, 2014

D. Lopez
O. Gonzalez de Dios
Telefonica I+D
Q. Wu
D. Dhody
Huawei
October 20, 2013

Secure Transport for PCEP
draft-lopez-pce-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 (TLS) and the TCP Authentication Option (TCP-AO) 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 April 23, 2014.

Copyright Notice

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

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of

publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Table of Contents

1. Introduction	3
2. Applying PCEPS	3
2.1. TCP ports	4
2.2. TLS Connection Establishment	4
2.3. TCP-AO Application	6
2.4. Peer Identity	6
2.5. Connection Establishment Failure	7
3. Discovery Mechanisms	7
3.1. DANE Applicability	8
4. Backward Compatibility	8
5. IANA Considerations	8
6. Security Considerations	8
7. Acknowledgements	9
8. References	9
8.1. Normative References	9
8.2. Informative References	10
Authors' Addresses	10

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]).

To further strengthen security mechanisms, the optional usage of the TCP Authentication Option (TCP-AO) [RFC5925] is introduced, and recommended especially in the case of long-lived connections.

This document describes a security container for the transport of PCEP requests and replies, and therefore it will not interfere with the protocol flexibility and extensibility.

This document describes how to apply TLS and TCP-AO in securing PCE interactions, including the TLS handshake mechanisms, the TLS methods for peer authentication, the applicable TLS ciphersuites for data exchange, the TCP-AO MKT establishment, and the handling of erros in the security checks. In the rest of the document we will refer to this usage of TLS and TCP-AO to provide a secure transport for PCEP as "PCEPS".

2. Applying PCEPS

2.1. TCP ports

The default destination port number for PCEPS is TCP/XXXX.

NOTE: This port has to be agreed and registered as PCEPS with IANA.

2.2. TLS Connection Establishment

PCEPS has no notion of negotiating TLS in an established connection. PCEP peers MAY either discover that the other PCEP endpoint supports PCEPS or can be preconfigured to use PCEPS for a given peer (see section Section 3 for more details). 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.

To support TLS re-negotiation both peers MUST support the mechanism described in [RFC5746]. Any attempt of initiate a TLS handshake to establish new cryptographic parameters not aligned with [RFC5746] SHALL be considered a TLS negotiation failure.

NOTE: We have to consider potential interactions between TLS re-negotiation and TCP-AO MKT

2.3. TCP-AO Application

PCEPS implementations MAY in addition apply the mechanisms described by the TCP Authentication Option (TCP-AO, described in [RFC5925]) to provide an additional level of protection with respect to attacks specifically addressed to forging the TCP connection underpinning TLS. TCP-AO is fully compatible with and deemed as complementary to TLS, so its usage is to be considered as a security enhancement whenever any of the PCEPS peers require it.

Implementations including support for TCP-AO MUST provide mechanisms to configure the requirements to use TCP-AO, as well as the association of a TCP-AO Master Key Tuple (MKT) with a particular peer. Whether these mechanisms are provided by the administrative interface or rely on the TLS handshake according to procedures similar to those described in [RFC5216] and [RFC5705] is outside the scope of this document.

2.4. 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

In addition, a PCC MAY apply the procedures described in [RFC6698] (DANE) to verify its peer identity when using DNS discovery. See section Section 3.1 for further details.

2.5. Connection Establishment Failure

In case the initial TLS negotiation, the peer identity check, or the optional TCP-AO MKT establishment fail according to the procedures listed in this document, the peer MUST immediately terminate the session. It SHOULD follow the procedure listed in [RFC5440] to retry session setup along with an exponential back-off session establishment retry procedure.

3. Discovery Mechanisms

A PCE can advertise its capability to support PCEPS using the IGP advertisement and discovery mechanism. 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. PCE capability bits are defined in [RFC5088].

NOTE: A new bit must be added here to advertise the PCEPS capability.

When DNS is used by a PCC willing to use PCEPS to locate an appropriate PCE [I-D.wu-pce-dns-pce-discovery], the PCC as initiating entity chooses at least one of the returned FQDNs to resolve, which it does by performing DNS "A" or "AAAA" lookups on the FDQN. This will eventually result in an IPv4 or IPv6 address. The PCC SHALL use the IP address(es) from the successfully resolved FDQN (with the corresponding port number returned by the DNS SRV lookup) as the connection address(es) for the receiving entity.

If the PCC fails to connect using an IP address but the "A" or "AAAA" lookups returned more than one IP address, then the PCC SHOULD use the next resolved IP address for that FDQN as the connection address.

If the PCC fails to connect using all resolved IP addresses for a given FQDN, then it SHOULD repeat the process of resolution and connection for the next FQDN returned by the SRV lookup based on the priority and weight.

If the PCC receives a response to its SRV query but it is not able to establish a PCEPS connection using the data received in the response, as initiating entity it MAY fall back to lookup a PCE that uses TCP as transport.

3.1. DANE Applicability

DANE [RFC6698] defines a secure method to associate the certificate that is obtained from a TLS server with a domain name using DNS, i.e., using the TLSA DNS resource record (RR) to associate a TLS server certificate or public key with the domain name where the record is found, thus forming a "TLSA certificate association". The DNS information needs to be protected by DNSSEC. A PCC willing to apply DANE to verify server identity MUST conform to the rules defined in section 4 of [RFC6698].

4. Backward Compatibility

Since the procedure described in this document describes a security container for the transport of PCEP requests and replies carried on a newly allocated TCP port there will be no impact on the base PCEP and/or any further extensions.

5. IANA Considerations

NOTE: PCEPS has to be registered as TCP port XXXX.

No new PCEP messages or other objects are defined.

6. 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.

7. Acknowledgements

This specification relies on the analysis and profiling of TLS included in [RFC6614].

8. References

8.1. 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.
- [RFC5746] Rescorla, E., Ray, M., Dispensa, S., and N. Oskov, "Transport Layer Security (TLS) Renegotiation Indication Extension", RFC 5746, February 2010.
- [RFC5925] Touch, J., Mankin, A., and R. Bonica, "The TCP Authentication Option", RFC 5925, June 2010.

- [RFC6066] Eastlake, D., "Transport Layer Security (TLS) Extensions: Extension Definitions", RFC 6066, January 2011.
- [RFC6698] Hoffman, P. and J. Schlyter, "The DNS-Based Authentication of Named Entities (DANE) Transport Layer Security (TLS) Protocol: TLSA", RFC 6698, August 2012.

8.2. Informative References

- [I-D.wu-pce-dns-pce-discovery]
Wu, W., Dhody, D., King, D., and D. Lopez, "Path Computation Element (PCE) Discovery using Domain Name System(DNS)", draft-wu-pce-dns-pce-discovery-03 (work in progress), October 2013.
- [RFC5216] Simon, D., Aboba, B., and R. Hurst, "The EAP-TLS Authentication Protocol", RFC 5216, March 2008.
- [RFC5705] Rescorla, E., "Keying Material Exporters for Transport Layer Security (TLS)", RFC 5705, March 2010.
- [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

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

Email: sunseawq@huawei.com

Dhruv Dhody
Huawei
-
Bangalore,
India

Phone: +91-9845062422
Email: sunseawq@huawei.com

PCE Working Group
Internet-Draft
Intended status: Standards Track
Expires: April 11, 2014

E. Crabbe
Google, Inc.
J. Medved
Cisco Systems, Inc.
I. Minei
Juniper Networks, Inc.
R. Varga
Pantheon Technologies SRO
X. Zhang
D. Dhody
Huawei Technologies
October 8, 2013

Optimizations of State Synchronization Procedures for Stateful PCE
draft-minei-pce-stateful-sync-optimizations-00

Abstract

A stateful Path Computation Element (PCE) has access to not only the information carried by the network's IGP, but also to the set of active paths and their reserved resources for its computations. The additional state allows the PCE to compute constrained paths while considering individual LSPs and their interactions. This requires reliable state synchronization mechanisms between the PCE and the network, PCE and path computation clients (PCCs), and between cooperating PCEs. The basic mechanism for state synchronization is part of the Stateful PCE specification. This draft specifies optimizations related to state synchronization procedures.

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 April 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	4
2. Terminology	4
3. State synchronization avoidance	4
3.1. Motivation	4
3.2. State Synchronization Avoidance procedures	5
3.3. LSP State Database Version Number TLV	9
3.3.1. Use of the LSP-DB-VERSION TLV in the OPEN object	10
3.3.2. Use of the LSP-DB-VERSION TLV in the LSP object	10
3.4. Speaker Entity Identifier TLV	10
4. PCE-triggered State Synchronization	11
4.1. Motivation	11
4.2. PCE-triggered State Synchronization Procedures	11
5. Incremental State Synchronization	12
5.1. Motivation	12
5.2. Incremental synchronization procedures	14
6. Advertising support of the synchronization optimizations	16
7. IANA Considerations	17
7.1. PCEP-Error Object	17
7.2. PCEP TLV Type Indicators	17
7.3. STATEFUL-PCE-CAPABILITY TLV	17
8. Security Considerations	18
9. Acknowledgements	18
10. Contributors	18
11. Normative References	18
Authors' Addresses	19

1. Introduction

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

[I-D.ietf-pce-stateful-pce] describes a set of extensions to PCEP to provide stateful control. A stateful PCE has access to not only the information carried by the network's IGP, but also to the set of active paths and their reserved resources for its computations. The additional state allows the PCE to compute constrained paths while considering individual LSPs and their interactions. This requires reliable state synchronization mechanisms between the PCE and the network, PCE and PCC, and between cooperating PCEs.

[I-D.ietf-pce-stateful-pce] describes the basic mechanism for state synchronization. This draft specifies optimizations for state synchronization.

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, Delegation Timeout Interval, LSP State Report, LSP Update Request, LSP Priority, LSP State Database, Revocation.

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

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

3. State synchronization avoidance

3.1. Motivation

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]). [I-D.ietf-pce-stateful-pce] describes the basic mechanism for state synchronization.

State synchronization is not always necessary following a PCEP

session restart. If the state of both PCEP peers did not change, the synchronization phase may be skipped. This can result in significant savings in both control-plane data exchanged and the time it takes for the session to become fully operational.

3.2. State Synchronization Avoidance procedures

State Synchronization MAY be skipped following a PCEP session restart if the state of both PCEP peers did not change during the period prior to session re-initialization. To be able to make this determination, state must be exchanged and maintained by both PCE and PCC during normal operation. This is accomplished by keeping track of the changes to the LSP State Database, using a version tracking field called the LSP State Database Version Number.

The LSP State Database Version Number is an unsigned 64-bit value that MUST be incremented by 1 for each successive change in the LSP state database. The LSP State Database Version Number MUST start at 1 and may wrap around. Values 0 and 0xFFFFFFFFFFFFFFFF are reserved. The PCC is the owner of the LSP State Database Version Number, which is incremented each time a change is made to the PCC's local LSP State Database. Operations that trigger a change to the local LSP State database include a change in the LSP operational state, delegation of an LSP, removal or addition of an LSP or change in any of the LSP attributes that would trigger a report to the PCE. When State Synchronization avoidance is enabled on a PCEP session, a PCC includes the LSP-DB-VERSION TLV in the LSP Object on each LSP State Report. The LSP-DB-VERSION TLV contains a PCC's LSP State Database version.

State Synchronization Avoidance is advertised on a PCEP session during session startup using the INCLUDE-DB-VERSION bit in the capabilities TLV (see Section 6). The peer may move in the network, either physically or logically, which may cause its connectivity details and transport-level identity (such as IP address) to change. To ensure that a PCEP peer can recognize a previously connected peer even in face of such mobility, each PCEP peer includes the SPEAKER-ENTITY-ID TLV described in Section 3.4 in the OPEN message.

If both PCEP speakers set the INCLUDE-DB-VERSION Flag in the OPEN object's STATEFUL-PCE-CAPABILITY TLV to 1, the PCC will include the LSP-DB-VERSION TLV in each LSP Object. The TLV will contain the PCC's latest LSP State Database Version Number.

If a PCE's LSP State Database survived the restart of a PCEP session, the PCE will include the LSP-DB-VERSION TLV in its OPEN object, and the TLV will contain the last LSP State Database Version Number received on an LSP State Report from the PCC in a previous PCEP

session. If a PCC's LSP State Database survived the restart of a PCEP session, the PCC will include the LSP-DB-VERSION TLV in its OPEN object and the TLV will contain the latest LSP State Database Version Number sent on an LSP State Report from the PCC in the previous PCEP session. If a PCEP Speaker's LSP State Database did not survive the restart of a PCEP session, the PCEP Speaker MUST NOT include the LSP-DB-VERSION TLV in the OPEN Object.

If both PCEP Speakers include the LSP-DB-VERSION TLV in the OPEN Object and the TLV values match, the PCC MAY skip State Synchronization. Otherwise, the PCC MUST perform State Synchronization. If the PCC attempts to skip State Synchronization (i.e. the SYNC Flag = 0 on the first LSP State Report from the PCC), the PCE MUST send back a PCErrror with Error-type 20 Error-value 2 'LSP Database version mismatch', and close the PCEP session.

If state synchronization is required, then prior to completing the Initialization phase, the PCE MUST mark any LSPs in the LSP database that were previously reported by the PCC as stale. When the PCC reports an LSP during state synchronization, if the LSP already exists in the LSP database, the PCE MUST update the LSP database and clear the stale marker from the LSP. When it has finished state synchronization, the PCC MUST immediately send an end of synchronization marker. The end of synchronization marker is a PCRpt message with an LSP object containing a PLSP-ID of 0 and with the SYNC flag set to 0 ([I-D.ietf-pce-stateful-pce]). The LSP-DB-VERSION TLV MUST be included and contain the PCC's latest LSP State Database Version Number. On receiving this state report, the PCE MUST purge any LSPs from the LSP database that are still marked as stale.

Note that a PCE/PCC MAY force State Synchronization by not including the LSP-DB-VERSION TLV in its OPEN object.

Figure 1 shows an example sequence where State Synchronization is skipped. In the figure, IDB stands for INCLUDE-DB-VERSION.

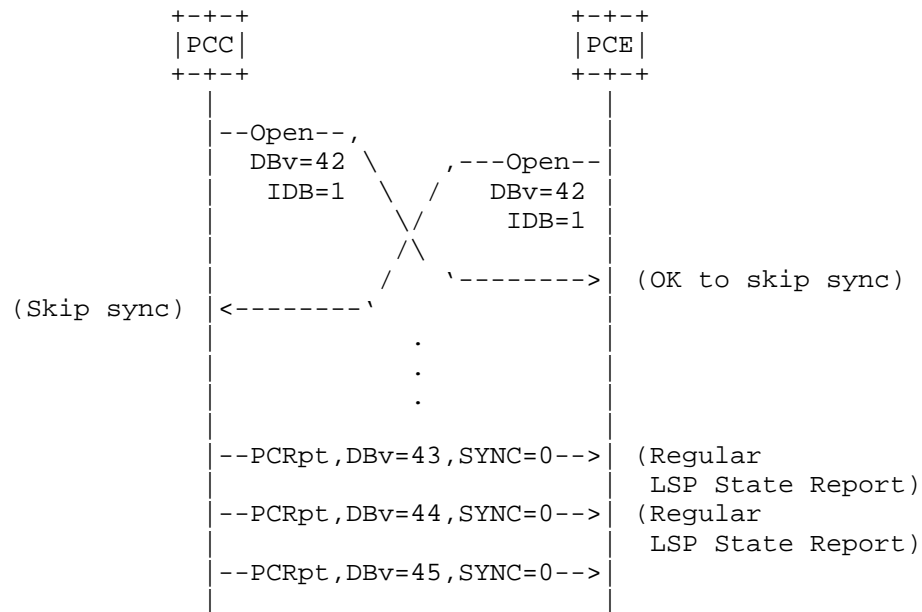


Figure 1: State Synchronization skipped

Figure 2 shows an example sequence where State Synchronization is performed due to LSP State Database version mismatch during the PCEP session setup. Note that the same State Synchronization sequence would happen if either the PCC or the PCE would not include the LSP-DB-VERSION TLV in their respective Open messages. In the figure, IDB stands for INCLUDE-DB-VERSION.

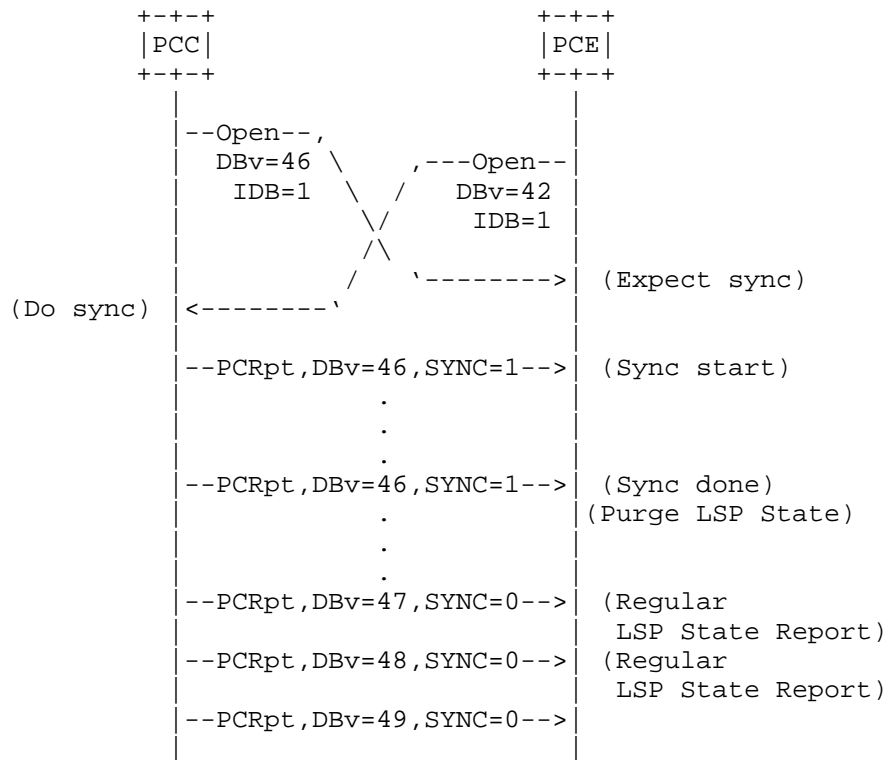


Figure 2: State Synchronization performed

Figure 3 shows an example sequence where State Synchronization is skipped, but because one or both PCEP Speakers set the INCLUDE-DB-VERSION Flag to 0, the PCC does not send LSP-DB-VERSION TLVs to the PCE. If the current PCEP session restarts, the PCEP Speakers will have to perform State Synchronization, since the PCE will not know the PCC's latest LSP State Database Version Number. In the figure IDB stands for INCLUDE-DB-VERSION.

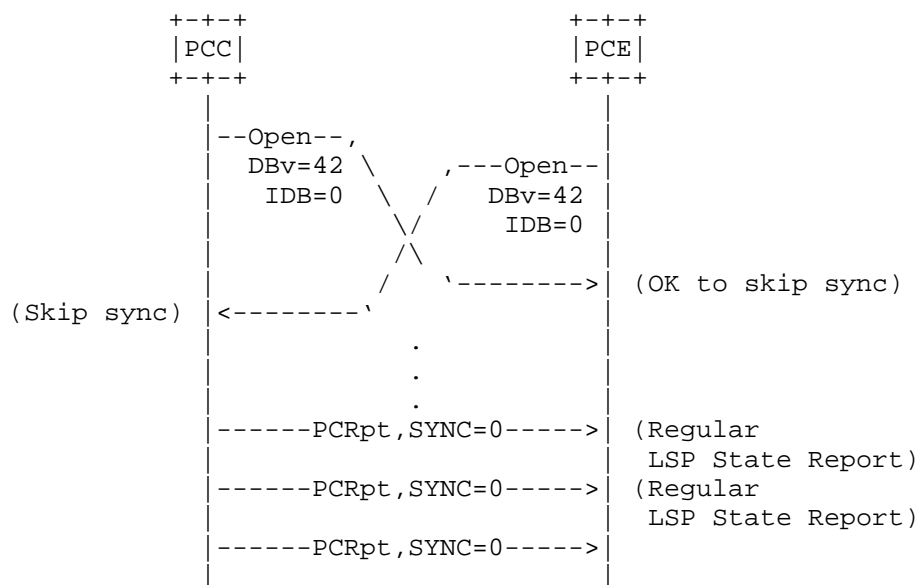


Figure 3: State Synchronization skipped, no LSP-DB-VERSION TLVs sent from PCC

3.3. LSP State Database Version Number TLV

The LSP State Database Version Number (LSP-DB-VERSION) TLV is an optional TLV that MAY be included in the OPEN object and the LSP object.

The format of the LSP-DB-VERSION TLV is shown in the following figure:

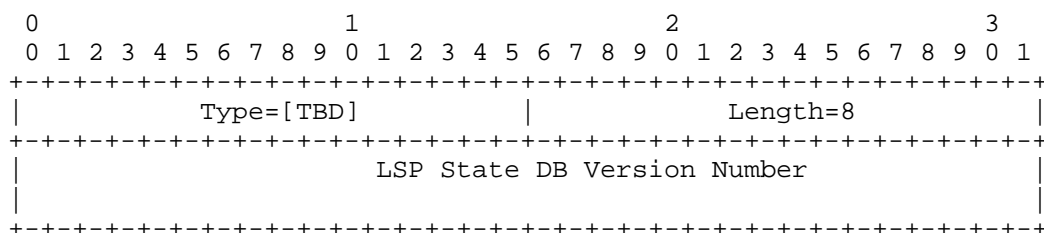


Figure 4: LSP-DB-VERSION TLV format

The type of the TLV is [TBD] and it has a fixed length of 8 octets. The value contains a 64-bit unsigned integer.

3.3.1. Use of the LSP-DB-VERSION TLV in the OPEN object

The LSP-DB-VERSION TLV is included as an optional TLV in the OPEN object when a PCEP Speaker wishes to determine if State Synchronization can be skipped when a PCEP session is restarted. If sent from a PCE, the TLV contains the local LSP State Database Version Number from the last valid LSP State Report received from a PCC. If sent from a PCC, the TLV contains the PCC's local LSP State Database Version Number, which is incremented each time the LSP State Database is updated.

3.3.2. Use of the LSP-DB-VERSION TLV in the LSP object

The LSP-DB-VERSION TLV can be included as an optional TLV in the LSP object.

If State Synchronization Avoidance has been enabled on a PCEP session (as described in Section 3.2), a PCC MUST include the LSP-DB-VERSION TLV in each LSP Object sent out on the session. If the TLV is missing, the PCE will generate an error with error-type 6 (mandatory object missing) and Error Value 12 (LSP-DB-VERSION TLV missing) and close the session. If State Synchronization Avoidance has not been enabled on a PCEP session, the PCC SHOULD NOT include the LSP-DB-VERSION TLV in the LSP Object and the PCE SHOULD ignore it were it to receive one.

Since a PCE does not make changes to the LSP State Database Version Number, a PCC should never encounter this TLV in a message from the PCE (other than the OPEN message). A PCC SHOULD ignore the LSP-DB-VERSION TLV, were it to receive one from a PCE.

If State Synchronization Avoidance is enabled, a PCC MUST increment its LSP State Database Version Number when the 'Redelegation Timeout Interval' timer expires (see [I-D.ietf-pce-stateful-pce] for the use of the Redelegation Timeout Interval).

3.4. Speaker Entity Identifier TLV

SPEAKER-ENTITY-ID is an optional TLV that MAY be included in the OPEN Object when a PCEP Speaker wishes to determine if State Synchronization can be skipped when a PCEP session is restarted. It contains a unique identifier for the node that does not change during the life time of the PCEP Speaker. It identifies the PCEP Speaker to its peers if the Speaker's IP address changed.

The format of the SPEAKER-ENTITY-ID TLV is shown in the following figure:

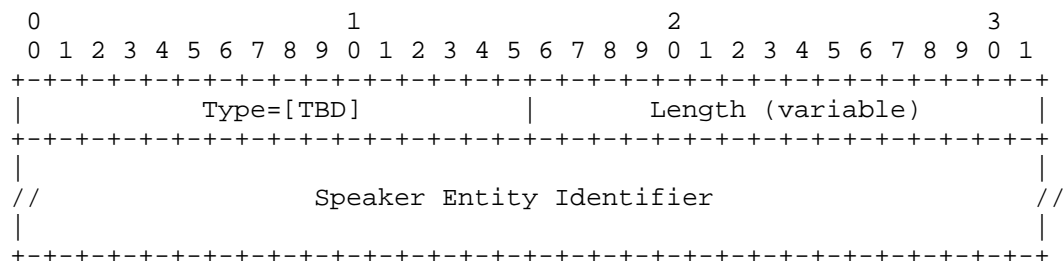


Figure 5: SPEAKER-ENTITY-ID TLV format

The type of the TLV is [TBD] and it has a variable length, which MUST be greater than 0. The value contains the entity identifier of the speaker transmitting this TLV. This identifier is required to be unique within its scope of visibility, which is usually limited to a single domain. It MAY be configured by the operator. Alternatively it can be derived automatically from a suitably-stable unique identifier, such as a MAC address, serial number, Traffic Engineering Router ID, or similar. In the case of inter-domain connections, the speaker SHOULD prefix its usual identifier with the domain identifier of its residence, such as Autonomous System number, IGP area identifier, or similar.

The relationship between this identifier and entities in the Traffic Engineering database is intentionally left undefined.

From a manageability point of view, a PCE or PCC implementation SHOULD allow the operator to configure a SPEAKER-ENTITY-ID.

4. PCE-triggered State Synchronization

4.1. Motivation

The accuracy of the computations performed by the PCE is tied to the accuracy of the view the PCE has on the state of the LSPs. Therefore, it can be beneficial to be able to resynchronize this state even after the session has established. The PCE may use this approach to continuously sanity check its state against the network, or to recover from error conditions without having to tear down sessions.

4.2. PCE-triggered State Synchronization Procedures

Support of PCE-triggered state synchronization is advertised on a PCEP session during session startup using the TRIGGERED-SYNC (T) bit in the capabilities TLV. The PCE can choose to resynchronize its

entire LSP database, or a single LSP.

To trigger resynchronization for an LSP, the PCE MUST first mark the LSP as stale and then send a PCUpd for it, with the SYNC flag set to 1. The PCE SHOULD NOT include any parameter updates for the LSP, and the PCC SHOULD ignore such updates if the SYNC flag is set. The PCC MUST reply with a PCRpt and SHOULD include the SRP-ID-number of the PCUpd that triggered the report.

The PCE can also trigger resynchronization of the entire LSP database. The PCE MUST first mark any LSPs in the LSP database that were previously reported by the PCC as stale and then send a PCUpd for an LSP object containing a PLSP-ID of 0 and with the SYNC flag set to 1. This PCUpd message is the trigger for the PCC to enter the synchronization phase as described in [I-D.ietf-pce-stateful-pce] and start sending PCRpt messages. After the receipt of the end-of-synchronization marker, the PCE will purge LSPs which were not refreshed. The SRP-ID-number of the PCUpd that triggered the report SHOULD be included in each of the PCRpt messages.

If the TRIGGERED-SYNC capability was not advertised and the PCC receives a PCUpd with the SYNC flag set to 1, it MUST send a PCErr with the SRP-ID-number of the PCUpd, error-type 20 and error-value 4.(see Section 7.1)

5. Incremental State Synchronization

[I-D.ietf-pce-stateful-pce] describes LSP state synchronization mechanism between PCCs and PCEs for a stateful PCE. After PCEP session set up, PCC compares the LSP State Database version with the PCE as described in Section 3. 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 section 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 during incremental synchronization.

5.1. 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 6 shows an example of LSP state

synchronization.

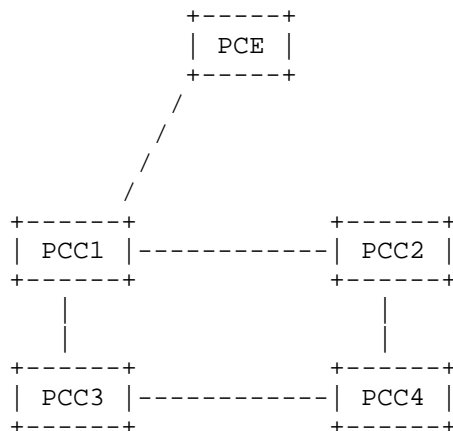


Figure 6: Topology Example

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 section, 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 trigger the LSP synchronization (similar to Section 4).

PCEP extensions for stateful PCEs to perform LSP synchronization SHOULD allow:

- o 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.

- o A stateful PCE to control the timing of PCC synchronizing its LSP state with the PCE during incremental synchronisation.

5.2. Incremental synchronization procedures

[I-D.ietf-pce-stateful-pce] describes state synchronization and Section 3 describes state synchronization avoidance by using LSP-DB-VERSION TLV in its OPEN object. This section 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.

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 7 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 [I-D.ietf-pce-stateful-pce].

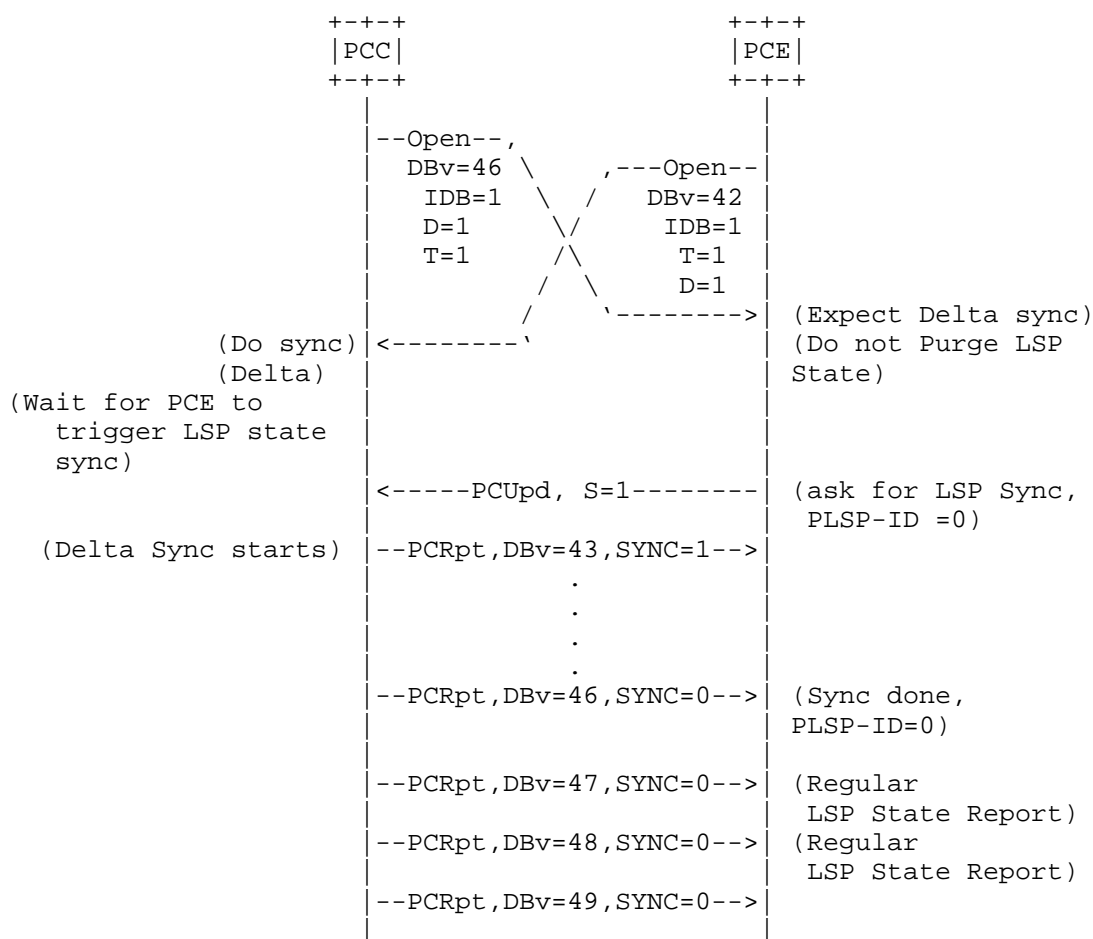


Figure 7: Incremental synchronization procedures

A stateful PCE MAY choose to control the LSP-DB synchronization process. To allow PCE to do so, PCEP speakers MUST set T bit to 1 to indicate this as described in Section 4. 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. Note that the PCE should not mark the existing LSPs as stale for incremental state synchronisation

procedure.

As per Section 3, 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 7, 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.

6. Advertising support of the synchronization optimizations

Support for each of the optimizations described in this document requires advertising support of the capability at session establishment time.

New flags are defined for the STATEFUL-PCE-CAPABILITY TLV defined in [I-D.ietf-pce-stateful-pce]. Its format is shown in the following figure:

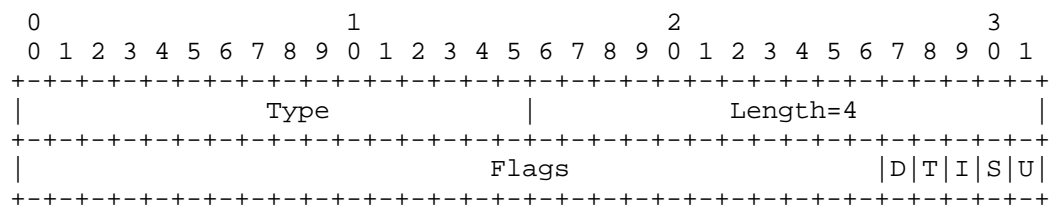


Figure 8: STATEFUL-PCE-CAPABILITY TLV format

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

U (LSP-UPDATE-CAPABILITY - 1 bit): defined in [I-D.ietf-pce-stateful-pce]

S (INCLUDE-DB-VERSION - 1 bit): if set to 1 by both PCEP Speakers, the PCC will include the LSP-DB-VERSION TLV in each LSP Object.

I (LSP-INSTANTIATION-CAPABILITY - 1 bit): defined in [I-D.crabbe-pce-pce-initiated-lsp]

T (TRIGGERED-SYNC - 1 bit): if set to 1 by both PCEP Speakers, the PCE can trigger synchronization of LSPs at any point in the life of the session. The flag must be advertised by both PCC and PCE for PCUpd messages with the SYNC flag set to be allowed on a PCEP session.

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.

7. IANA Considerations

This document requests IANA actions to allocate code points for the protocol elements defined in this document. Values shown here are suggested for use by IANA.

7.1. PCEP-Error Object

This document defines new Error-Value values for the LSP State synchronization error defined in [I-D.ietf-pce-stateful-pce].

Error-Type	Meaning
6	Mandatory Object missing Error-value=12: LSP-DB-VERSION TLV missing
20	LSP State synchronization error. Error-value=2: LSP Database version mismatch. Error-value=3: The LSP-DB-VERSION TLV Missing when State Synchronization Avoidance enabled. Error-value=4: Attempt to trigger a synchronization when the TRIGGERED-SYNC capability has not been advertised.

7.2. PCEP TLV Type Indicators

This document defines the following new PCEP TLVs:

Value	Meaning	Reference
23	LSP-DB-VERSION	This document
24	SPEAKER-ENTITY-ID	This document

7.3. STATEFUL-PCE-CAPABILITY TLV

The following values are defined in this document for the Flags field in the STATEFUL-PCE-CAPABILITY-TLV in the OPEN object:

Bit	Description	Reference
28	DELTA-LSP-SYNC-CAPABILITY	This document
29	TRIGGERED-SYNC	This document
30	INCLUDE-DB-VERSION	This document

8. Security Considerations

The security considerations listed in [I-D.ietf-pce-stateful-pce] apply to this document as well.

9. Acknowledgements

We would like to thank Young Lee for his contributions.

10. Contributors

Gang Xie
Huawei Technologies
F3-5-B R&D Center, Huawei Industrial Base, Bantian, Longgang District
Shenzhen, Guangdong 518129
P.R.China
Email: xiegang09@huawei.com

11. 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-06 (work in progress), August 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.

[RFC5511] Farrel, A., "Routing Backus-Naur Form (RBNF): A Syntax
Used to Form Encoding Rules in Various Routing Protocol
Specifications", RFC 5511, April 2009.

Authors' Addresses

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

Email: edc@google.com

Jan Medved
Cisco Systems, Inc.
170 West Tasman Dr.
San Jose, CA 95134
US

Email: jmedved@cisco.com

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

Email: ina@juniper.net

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

Email: robert.varga@pantheon.sk

Xian Zhang
Huawei Technologies
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 Technologies
Leela Palace
Bangalore, Karnataka 560008
INDIA

Email: dhruv.ietf@gmail.com

PCE Working Group
Internet-Draft
Intended status: Standards Track
Expires: April 19, 2014

S. Sivabalan
J. Medved
Cisco Systems, Inc.
I. Minei
Juniper Networks, Inc.
R. Varga
Pantheon Technologies SRO
E. Crabbe
Google, Inc.
October 16, 2013

Conveying path setup type in PCEP messages
draft-sivabalan-pce-lsp-setup-type-01.txt

Abstract

A Path Computation Element can compute traffic engineering paths (TE paths) through a network that are subject to various constraints. Currently, TE paths are label switched paths (LSPs) which are set up using the RSVP-TE signaling protocol. However, other TE path setup methods are possible within the PCE architecture. This document proposes an extension to PCEP to allow support for different path setup methods over a given PCEP session.

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 April 19, 2014.

Copyright Notice

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

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

Table of Contents

1. Introduction	2
2. Terminology	3
3. Path Setup Type TLV	3
4. Operation	4
5. Security Considerations	5
6. IANA Considerations	5
7. Acknowledgements	6
8. Normative References	6
Authors' Addresses	6

1. Introduction

[RFC5440] describes the 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 PCC requests a path subject to various constraints and optimization criteria from a PCE. The PCE responds to the PCC with a hop-by-hop path in an Explicit Route Object (ERO). The PCC uses the ERO to set up the path in the network.

[I-D.ietf-pce-stateful-pce] specifies extensions to PCEP that allow a PCC to delegate its LSPs to a PCE. The PCE can then update the state of LSPs delegated to it. In particular, the PCE may modify the path of an LSP by sending a new ERO. The PCC uses this ERO to re-route the LSP in a make-before-break fashion.

[I-D.crabbe-pce-pce-initiated-lsp] specifies a mechanism allowing a PCE to dynamically instantiate an LSP on a PCC by sending the ERO and characteristics of the LSP. The PCC signals the LSP using the ERO and other attributes sent by the PCE.

So far, the PCEP protocol and its extensions implicitly assume that the TE paths are label switched, and are established via the RSVP-TE protocol. However, other methods of LSP setup are not precluded. When a new path setup method (other than RSVP-TE) is introduced for setting up a path, a new capability TLV pertaining to the new path setup method MAY be advertised when the PCEP session is established. Such capability TLV MUST be defined in the specification of the new path setup type. When multiple path setup methods are deployed in a network, a given PCEP session may have to simultaneously support more than one path setup types. In this case, the intended path setup method needs to be either explicitly indicated or implied in the appropriate PCEP messages (when necessary) so that both the PCC and the PCE can take the necessary steps to set up the path. This document introduces a generic TLV called "PATH-SETUP-TYPE TLV" and specifies the base procedures to facilitate such operational model.

2. Terminology

The following terminologies are used in this document:

ERO: Explicit Route Object.
 LSR: Label Switching Router.
 PCC: Path Computation Client.
 PCE: Path Computation Element
 PCEP: Path Computation Element Protocol.
 TLV: Type, Length, and Value.

3. Path Setup Type TLV

When a PCEP session is used to set up TE paths using different methods, the corresponding PCE and PCC must be aware of the path setup method used. That means, a PCE must be able to specify paths in the correct format and a PCC must be able take control and take forwarding plane actions appropriate to the path setup type.

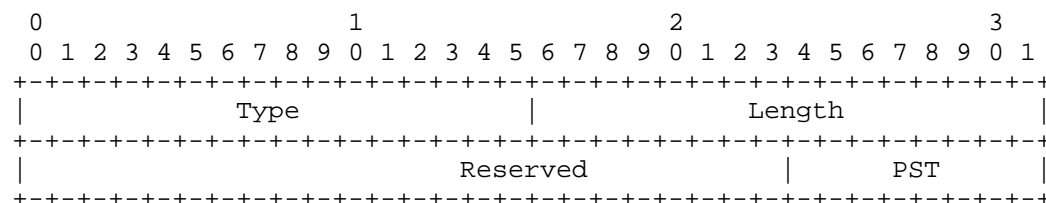


Figure 1: PATH-SETUP-TYPE TLV

PATH-SETUP-TYPE TLV is an optional TLV associated with the RP ([RFC5440]) and the SRP ([I-D.ietf-pce-stateful-pce]) objects. Its format is shown in the above figure. The type of the TLV is to be

defined by IANA. The one octet value contains the Path Setup Type (PST). This document specifies the following PST value:

- o PST = 0: Path is setup via RSVP-TE signaling protocol(default).

The absence of the PATH-SETUP-TYPE TLV is equivalent to an PATH-SETUP-TYPE TLV with an PST value of 0. It is recommended to omit the TLV in the default case. If the RP or SRP object contains more than one PATH-SETUP-TYPE TLVs, only the first TLV MUST be processed and the rest MUST be ignored.

If a PCEP speaker does not recognize the PATH-SETUP-TYPE TLV, it MUST ignore the TLV in accordance with ([RFC5440]). If a PCEP speaker recognizes the TLV but does not support the TLV, it MUST send PCErr with Error-Type = 2 (Capability not supported).

4. Operation

When requesting a path from a PCE using a PCReq message ([RFC5440]), a PCC MAY include the PATH-SETUP-TYPE TLV in the RP object. If the PCE is capable of expressing the path in a format appropriate to the setup method used, it MUST use the appropriate ERO format in the PCRep message. If the path setup type cannot be inferred from the ERO or any other object or TLV in the PCRep message, PATH-SETUP-TYPE TLV may be included in the RP object of the PCRep message. Regardless of whether PATH-SETUP-TYPE TLV is used or not, if the PCE does not support the intended path setup type it MUST send PCErr with Error-Type = TBD (Traffic engineering path setup error) (recommended value is 21) and Error-Value = 1 (Unsupported path setup type) and close the PCEP session. If the path setup types corresponding to the PCReq and PCRep messages do not match, the PCC MUST send a PCErr with Error-Type = 21 (Traffic engineering path setup error) and Error-Value = 2 (Mismatched path setup type) and close the PCEP session.

In the case of stateful PCE, if the path setup type cannot be unambiguously inferred from ERO or any other object or TLV, PATH-SETUP-TYPE TLV MAY be used in PCRpt and PCUpd messages. If PATH-SETUP-TYPE TLV is used in PCRpt message, the SRP object MUST be present even in cases when the SRP-ID-number is the reserved value of 0x00000000. Regardless of whether PATH-SETUP-TYPE TLV is used or not, if a PCRpt message is triggered due to a PCUpd message (in this case SRP-ID-number is not equal to 0x00000000), the path setup types corresponding to the PCRpt and PCUpd messages should match. Otherwise, the PCE MUST send PCErr with Error-Type = 21 (Traffic engineering path setup error) and Error-Value = 2 (Mismatched path setup type) and close the connection.

In the case of PCE initiated LSPs, a PCE MAY include PATH-SETUP-TYPE TLV in PCInitiate message if the message does not have any other means of indicating path setup type. If a PCC does not support the path setup type associated with the PCInitiate message, the PCC MUST send PCErr with Error-Type = 21 (Traffic engineering path setup error) and Error-Value = 1 (Unsupported path setup type) and close the PCEP session. Similarly, as mentioned above, if the path setup type cannot be unambiguously inferred from ERO or any other object or TLV, the PATH-SETUP-TYPE TLV MAY be included in PCRpt messages triggered by PCInitiate message. Regardless of whether PATH-SETUP-TYPE TLV is used or not, if a PCRpt message is triggered by a PCInitiate message, the path setup types corresponding to the PCRpt and the PCInitiate messages should match. Otherwise, the PCE MUST send PCErr message with Error-Type = 21 (Traffic engineering path setup error) and Error-Value = 2 (Mismatched path setup type). If PATH-SETUP-TYPE TLV is used in PCRpt message, SRP object MUST be included in PCRpt message even if SRP-ID-number is the reserved value of 0x00000000.

5. Security Considerations

No additional security measure is required.

6. IANA Considerations

IANA is requested to allocate a new TLV type (recommended value is TBD) for PATH-SETUP-TYPE TLV specified in this document.

This document requests that a registry is created to manage the value of the path Setup Type field in the PATH-SETUP-TYPE TLV.

Value	Description	Reference
0	Traffic engineering path is setup using RSVP signaling protocol	This document

Table 1

This document also defines a new Error-Type (recommended 21) and new Error-Values for the following new error conditions:

Error-Type	Meaning
21	Invalid traffic engineering path setup type
Error-value=1:	Unsupported path setup type
Error-value=2:	Mismatched path setup type

7. Acknowledgements

We like to thank Marek Zawodsky for valuable comments.

8. 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-03 (work in progress), October 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-07 (work in progress), October 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.

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
USA

Email: jmedved@cisco.com

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

Email: ina@juniper.net

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

Email: robert.vargad@pantheon.sk

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

Email: edc@google.com

PCE Working Group
Internet-Draft
Intended status: Standards Track
Expires: April 14, 2014

Q. Wu
D. Dhody
Huawei
D. King
Old Dog Consulting
D. Lopez
Telefonica I+D
October 11, 2013

Path Computation Element (PCE) Discovery using Domain Name System(DNS)
draft-wu-pce-dns-pce-discovery-03

Abstract

Discovery of the Path Computation Element (PCE) within an IGP area or routing domain is possible using OSPF [RFC5088] and IS-IS [RFC5089]. However, it has been established that in certain deployment scenarios PCEs may not wish, or be able to participate within the IGP process. In those scenarios, it is beneficial for the Path Computation Client (PCC) (or other PCE) 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 PCE type and capability discovery via DNS.

Status of this Memo

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

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

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

This Internet-Draft will expire on April 14, 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. Terminology	3
1.2. Requirements	3
2. Conventions used in this document	5
3. Motivation	6
3.1. Outside the Routing Domain	6
3.2. Discovery Mechanisms	7
3.2.1. Query-Response versus Advertisement	7
3.3. Network Address Translation Gateway	7
4. Additional Capabilities	8
4.1. Load Sharing of Path Computation Requests	8
5. Extended Naming Authority Pointer (NAPTR)Service Field Format	9
5.1. IETF Standards Track PCE Applications	10
6. Backwards Compatibility	11
7. Discovering a Path Computation Element	12
7.1. Determining the PCE Service and transport protocol	13
7.2. Determining the IP Address of the PCE	13
7.2.1. Examples	15
7.3. Determining path computation scope, capability, the PCE domains and Neighbor PCE domains	16
8. IANA Considerations	17
8.1. IETF PCE Application Service Tags	17
8.2. PCE Application Protocol Tags	17
9. Security Considerations	18
10. Acknowledgements	19
11. References	20
11.1. Normative References	20
11.2. Informative References	21
Authors' Addresses	23

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 PCE) 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.

This document describes PCE discovery via DNS. The mechanism with which DNS comes to know about the PCE and its capability is out of scope of this document.

1.1. Terminology

The following terminology is used in this document.

PCE-Domain: As per [RFC4655], any collection of network elements within a common sphere of address management or path computational responsibility. Examples of domains include Interior Gateway Protocol (IGP) areas and Autonomous Systems (ASs).

Domain-Name: An identification string that defines a realm of administrative autonomy, authority, or control on the Internet. Any name registered in the DNS is a domain name. DNS Domain names are used in various networking contexts and application-specific naming and addressing purposes. In general, a domain name represents an Internet Protocol (IP) resource. Examples of DNS domain name is "www.example.com" or "example.com"[RFC1035].

1.2. 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., 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;

These PCE discovery information allows PCCs to select appropriate PCEs:

This document specifies the procedures and extension to facilitate DNS-based PCE information discovery for specific use cases, and to complement existing IGP 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. Outside the Routing Domain

When the PCE is a router participating in the IGP, or even a server participating passively in the IGP, with all PCEP speakers in the same routing domain, a simple and efficient way to announce PCEs consists of using IGP flooding.

It has been identified that the existing PCE discovery mechanisms do not work in following scenarios:

Inter-AS: Per domain path computation mechanism [RFC5152] or Backward recursive path computation (BRPC) [RFC5441] MAY be used by cooperating PCEs to compute inter-domain path. In which case these cooperating PCEs should be known to other PCEs. In case of inter-AS where the PCEs do not participate in a common IGP, the existing IGP discovery mechanism cannot be used to discover inter-AS PCE.

Hierarchy of PCE: 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 [PCE-QUESTION]. [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 to child PCEs when they are not a part of the same routing domain.

BGP: [BGP-LS] describes a mechanism by which links state and traffic engineering information can be collected from networks and shared with external components using the BGP routing protocol. An external PCE MAY use this mechanism to populate its TED and not take part in the same IGP routing domain.

NMS/OSS: PCE MAY gain the knowledge of Topology information from some management system (e.g., NMS/OSS) and not take part in the same routing domain. Also note that in some case PCC may not be a router and instead be a management system like NMS and may not be able to discover PCE via IGP discovery.

3.2. Discovery Mechanisms

3.2.1. Query-Response versus Advertisement

Advertisement based PCE discovery using IGP methods [RFC5088] and [RFC5089] floods the PCE information to an area, a subset of areas or to a full routing domain. By the very nature of flooding and advertisements it generates unwanted traffic and may lead to unnecessary advertisement, especially when PCE information needs frequent changes.

DNS is a query-response based mechanism, a client (a PCC) can use DNS to discover a PCE only when it needs to compute a path and does not require any other node in the network to be involved.

In case of Intermittent PCEP session, where PCEP sessions are systematically open and closed for each PCEP request, a DNS-based query-response mechanism is more suitable. One may also utilize DNS-based load-balancing and recovery functions.

3.3. Network Address Translation Gateway

PCEP uses TCP as the transport mechanism between PCC and PCE, and PCE to PCE, communications [RFC5440]. To secure TCP connection that underlay PCEP sessions, Transport Layer Security (TLS) can be used besides using TCP-MD5 [RFC2385] and TCP-AUTH [RFC5295]. When PCC and PCE support TCP-MD5 or TCP-AUTH while NAT does not, TCP connection establishment fails. When NAT gateway is in presence, a TCP or TCP/TLS connection can be opened by Interactive Connectivity Establishment (ICE) [RFC5245] for the purpose of connectivity checks. However the TCP connection cannot be established in cases where one of the peers 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.

4. Additional Capabilities

4.1. Load Sharing of Path Computation Requests

Multiple PCEs can be present in a single network domain for redundancy. DNS supports inherent load balancing where multiple PCEs (with different IP addresses) are known in DNS for a single PCE server name and are hidden from the PCC.

In an IGP advertisement based PCE discovery, one learns of all the PCEs and it is the job of the PCC to do load-balancing.

A DNS-based load-balancing mechanism works well in case of Intermittent PCEP sessions and request are load-balanced among PCEs similar to HTTP request without any complexity at the client.

5. Extended Naming Authority Pointer (NAPTR)Service Field Format

The NAPTR service field format defined by the S-NAPTR DDDS application in [RFC3958] follows this Augmented Backus-Naur Form (ABNF) [RFC5234]:

```

service-parms = [ [app-service] *(":" app-protocol)]
app-service   = experimental-service / iana-registered-service
app-protocol  = experimental-protocol / iana-registered-protocol
experimental-service      = "x-" 1*30ALPHANUMSYM
experimental-protocol     = "x-" 1*30ALPHANUMSYM
iana-registered-service   = ALPHA *31ALPHANUMSYM
iana-registered-protocol  = ALPHA *31ALPHANUMSYM
ALPHA                    = %x41-5A / %x61-7A ; A-Z / a-z
DIGIT                    = %x30-39 ; 0-9
SYM                      = %x2B / %x2D / %x2E ; "+" / "-" / "."
ALPHANUMSYM              = ALPHA / DIGIT / SYM
; The app-service and app-protocol tags are limited to 32
; characters and must start with an alphabetic character.
; The service-parms are considered case-insensitive.

```

This specification refines the "iana-registered-service" tag definition for the discovery of PCE supporting a specific PCE application or capability as defined below.

```

iana-registered-service =/ pce-service
pce-service             = "pce+ap" appln-id
appln-id                = 1*10DIGIT
                        ; Application Identifier expressed as
                        ; a decimal integer without leading
                        ; zeros.

```

The appln-id element is the Application Identifier used to identify a specific PCE application. The PCE Application Identifier is a 32-bit unsigned integer, and values are allocated by IANA as defined in section 8.1.

This specification also refines the "iana-registered-protocol" tag definition for the discovery of PCE supporting a specific transport protocol as defined below.

```

iana-registered-protocol =/ pce-protocol
pce-protocol             = "pce." pce-transport
pce-transport            = "tcp" / "tls.tcp"

```

Similar to application protocol tags defined in the [RFC6408], the S-NAPTR application protocol tags defined by this specification MUST NOT be parsed in any way by the querying application or Resolver.

The delimiter (".") is present in the tag to improve readability and does not imply a structure or namespace of any kind. The choice of delimiter (".") for the application protocol tag follows the format of existing S-NAPTR application protocol tag registry entries, but this does not imply that it shares semantics with any other specifications that create registry entries with the same format.

The S-NAPTR application service and application protocol tags defined by this specification are unrelated to the IANA "Service Name and Transport Protocol Port Number Registry" (see [RFC6335]).

The maximum length of the NAPTR service field is 256 octets, including a one-octet length field (see Section 4.1 of [RFC3403] and Section 3.3 of [RFC1035]).

5.1. IETF Standards Track PCE Applications

A PCE Client MUST be capable of using the extended S-NAPTR application service tag for dynamic discovery of a PCE supporting Standards Track applications. Therefore, every IETF Standards Track PCE application MUST be associated with a "PCE-service" tag formatted as defined in this specification and allocated in accordance with IANA policy (see Section 8).

For example, a NAPTR service field value of:

```
'PCE+apl:pce.tcp'
```

means that the PCE in the SRV or A/AAAA record supports the Global Concurrent Optimization Application ('1') (See section 8.1) and the Transport Control Protocol (TCP) as the transport protocol (See section 8.2).

6. Backwards Compatibility

Domain Name System (DNS) administrators SHOULD also provision legacy NAPTR records [RFC3403] in order to guarantee backwards compatibility with legacy PCE that only support S-NAPTR DDDS application in [RFC3958]. If the DNS administrator provisions both extended S-NAPTR records as defined in this specification and legacy NAPTR records defined in [RFC3403], then the extended S-NAPTR records MUST have higher priority(e.g., lower order and/or preference values) than legacy NAPTR records.

7. Discovering a Path Computation Element

The extended-format NAPTR records provide a mapping from a domain to the SRV record or A/AAAA record for contacting a server supporting a specific transport protocol and PCE application. The resource record will contain an empty regular expression and a replacement value, which is the SRV record or the A/AAAA record for that particular transport protocol.

The assumption for this mechanism to work is that the DNS administrator of the queried domain has first provisioned the DNS with extended-format NAPTR entries.

When the PCC or other PCEs performs a NAPTR query for a server in a particular realm, the PCC or other PCEs has to know in advance the search path of the resolver, i.e., in which realm to look for a PCE, and in which Application Identifier it is interested.

The search path of the resolver can either be pre-configured, or discovered using Diameter, DHCP or other means. For example, the realm could be deduced from the Network Access Identifier (NAI) in the User-Name attribute-value pair (AVP) or extracted from the Destination-Realm AVP in Diameter [RFC6733].

When pre-configuration is used, PCE domain(e.g., AS200) can be added as "subdomains" of the first-level domain of the underlying service (e.g., AS200.example.com), which allows a NAPTR query for a server in a PCE domain associated with DNS domain-name.

When DHCP is used, it SHOULD know the domain-name of that realm and use DHCP to discover IP address of the PCE in that realm that provides path computation service along with some PCE location information useful to a PCC (or other PCE) for a 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 S-NAPTR lookups [RFC3958]. If neither the IP address nor other PCE location information can be discovered with the above procedure, the PCC (or other PCE) MAY request a domain search list, as described in [RFC3397] and [RFC3646], and use it as input to the DDDS application.

When the PCC (or other PCE) does not find valid domain-names using the mechanisms above, it MUST stop the attempt to discover any PCE.

The following procedures result in an IP address, PCE domain, neighboring PCE domain and PCE Computation Scope where the PCC (or other PCE) can contact the PCE that hosts the service it is looking for.

7.1. Determining the PCE Service and transport protocol

The PCC (or other PCE) should know the service identifier for the Path Computation service and associated transport protocol. The service identifier for the Path Computation service is defined as "PCE+apX" as specified in section 5, The PCE supporting "PCE" service MUST support TCP as transport, as described in [RFC5440].

The services relevant for the task of transport protocol selection are those with S-NAPTR service fields with values "PCE+apX:Y", where 'PCE+apX' is the service identifier defined in the previous paragraph, and ' Y' is the letter that corresponds to a transport protocol supported by the PCE. This document also establishes an IANA registry for mappings of S-NAPTR service name to transport protocol.

These NAPTR [RFC3958] records provide a mapping from a domain to the SRV [RFC2782] record for contacting a PCE with the specific transport protocol in the S-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 (or other PCE) MUST discard any service fields that identify a resolution service whose value is not valid. The S-NAPTR processing as described in [RFC3403] will result in the discovery of the most preferred PCE that is supported by the client, as well as an SRV record for the PCE.

7.2. Determining the IP Address of the PCE

If the returned NAPTR service fields contain entries formatted as "pce+apX:Y" where "X" indicates the Application Identifier and "Y" indicates the supported transport protocol(s), the target realm supports the extended format for NAPTR-based PCE discovery defined in this document.

- o If "X" contains the required Application Identifier and "Y" matches a supported transport protocol, the PCEP implementation resolves the "replacement" field entry to a target host using the lookup method appropriate for the "flags" field.
- o If "X" does not contain the required Application Identifier or "Y" does not match a supported transport protocol, the PCEP implementation abandons the peer discovery.

If the returned NAPTR service fields contain entries formatted as

"pce+apX" where "X" indicates the Application Identifier, the target realm supports the extended format for NAPTR-based PCE discovery defined in this document.

- o If "X" contains the required Application Identifier, the PCEP implementation resolves the "replacement" field entry to a target host using the lookup method appropriate for the "flags" field and attempts to connect using all supported transport protocols.
- o If "X" does not contain the required Application Identifier, the PCEP implementation abandons the PCE discovery.

If the returned NAPTR service fields contain entries formatted as "pce:X" where "X" indicates the supported transport protocol(s), the target realm supports PCEP but does not support the extended format for NAPTR-based PCE discovery defined in this document.

- o If "X" matches a supported transport protocol, the PCEP implementation resolves the "replacement" field entry to a target host using the lookup method appropriate for the "flags" field.

If the returned NAPTR service fields contain entries formatted as "pce", the target realm supports PCEP but does not support the extended format for NAPTR-based PCE discovery defined in this document. The PCEP implementation resolves the "replacement" field entry to a target host using the lookup method appropriate for the "flags" field and attempts to connect using TCP (in future it SHOULD attempt all supported transport Protocols) .

Note that the regexp field in the S-NAPTR example above is empty. The regexp field MUST NOT be used when discovering PCE, as its usage can be complex and error prone. Also, the discovery of the PCE does not require the flexibility provided by this field over a static target present in the TARGET field.

As the default behavior, the client is configured with the information about which transport protocol is used for a path computation service in a particular domain. The client 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_PCE._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.

7.2.1. Examples

As an example, consider a client that wishes to find PCED service in the as100.example.com domain. The client performs a S-NAPTR query for that domain, and the following NAPTR records are returned:

Order	Pref	Flags	Service	Regexp	Replacement
IN	NAPTR	50	50	"s" "pce:pce.tls.tcp"	" "
					_PCE._tcp.as100.example.com
IN	NAPTR	90	50	"s" "pce:pce.tcp"	" "
					_PCE._tcp.as100.example.com

This indicates that the domain does have a PCE providing Path Computation services over TCP, in that order of preference. If the client only supports TCP, TCP will be used, targeted to a host determined by an SRV lookup of _PCE._tcp.example.com. That lookup would return:

	;;	Priority	Weight	Port	Target
IN	SRV	0	1	XXXX	server1.as100.example.com
IN	SRV	0	2	XXXX	server2.as100.example.com

where XXXX represents the port number at which the service is reachable.

As an alternative example, a client wishes to discover a PCE in the ex2.example.com realm that supports the GCO application over TCP. The client performs a NAPTR query for that domain, and the following NAPTR records are returned:

```

;;      order pref flags service  regexp replacement
IN NAPTR 150  50  "a"   "pce:pce.tcp"  ""
        server1.ex2.example.com
IN NAPTR 150  50  "a"   "pce:pce.tls.tcp" ""
        server2.ex2.example.com
IN NAPTR 150  50  "a"   "pce+apl:pce.tcp" ""
        server1.ex2.example.com
IN NAPTR 150  50  "a"   "pce+apl:pce.tls.tcp" ""
        server2.ex2.example.com

```

This indicates that the server supports GCO(ID=1) over TCP and TLS/TCP via hosts server1.ex2.example.com and server2.ex2.example.com, respectively.

7.3. Determining path computation scope, capability, the PCE domains and Neighbor PCE domains

DNS servers MAY use DNS TXT record to give additional information about PCE service and add such TXT record to the additional information section (See section 4.1 of [RFC1035]) that are relevant to the answer and have the same authenticity as the data (Generally this will be made up of A and SRV records) in the answer section. The additional information may include path computation scope, capability, the PCE domains and Neighbor PCE domains associated with the PCE. If discovery of PCE supporting a specific PCE capability described in section 7.2 has already been performed, capability associated with the PCE does not need to be included in the additional information.

To store new types of information, the TXT record uses a structured format in its TXT-DATA field [RFC1035]. The format consists of the attribute name followed by the value of the attribute. The name and value are separated by an equals sign (=). The general syntax is defined in section 2 of [RFC1464] as follows:

```
<owner> <class> <ttl> TXT "<attribute name>=<attribute value>"
```

For example, the following TXT records contain attributes specified in this fashion:

```

ex2.example.com    IN    TXT    "path computation scope=inter-as"
ex2.example.com    IN    TXT    "capability=link constraint"
ex2.example.com    IN    TXT    "pce domain = as10"
ex2.example.com    IN    TXT    "neighbor pce domain= area1"

```

The PCC MAY inspect those Additional Information section in the DNS message and be capable of handling responses from nameservers that never fill in the Additional Information part of a response.

8. IANA Considerations

8.1. IETF PCE Application Service Tags

IANA specifies to create a new registry ' S-NAPTR application service tags' for existing IETF PCE applications.

Tag	PCE Application
pce+ap1	GCO [RFC5557]
pce+ap2	P2MP [RFC5671]
pce+ap3	Stateful [STATEFUL-PCE]
pce+ap4	GMPLS [RFC7025]
pce+ap5	Inter-AS[RFC5376]
pce+ap6	Inter-Area [RFC4927]
pce+ap7	Inter-layer [RFC6457]

Future IETF PCE applications MUST reserve the S-NAPTR application service tag corresponding to the allocated PCE Application ID as defined in Section 3.

8.2. PCE Application Protocol Tags

IANA has reserved the following S-NAPTR Application Protocol Tags for the PCE transport protocols in the "S-NAPTR Application Protocol Tag" registry created by [RFC3958].

Tag	Protocol
pce.tcp	TCP
pce.tls.tcp	TLS/TCP

Future PCE versions that introduce new transport protocols MUST reserve an appropriate S-NAPTR Application Protocol Tag in the "S-NAPTR Application Protocol Tag" registry created by [RFC3958].

9. Security Considerations

This document specifies an enhancement to the NAPTR service field format. The enhancement and modifications are based on the S-NAPTR, which is actually a simplification of the NAPTR, and therefore the same security considerations described in [RFC3958] are applicable to this document.

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 [RFC6781] 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].

However a malicious host doing S-NAPTR queries learns applications supported by PCEs in a certain realm faster, which might help the malicious host to scan potential targets for an attack more efficiently when some applications have known vulnerabilities.

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.

10. Acknowledgements

The author would like to thank Claire Bi, Ning Kong, Liang Xia, Stephane Bortzmeyer, Yi Yang, Ted Lemon, Adrian Farrel and Stuart Cheshire for their review and comments that help improvement to this document.

11. References

11.1. Normative References

- [RFC1035] Mockapetris, P., "DOMAIN NAMES - IMPLEMENTATION AND SPECIFICATION", RFC 1035, November 1987.
- [RFC1464] Rosenbaum, R., "Using the Domain Name System To Store Arbitrary String Attributes", RFC 1464, May 1993.
- [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.
- [RFC3958] Daigle, D. and A. Newton, "Domain-Based Application Service Location Using SRV RRs and the Dynamic Delegation Discovery Service (DDDS)", RFC 3958, January 2005.
- [RFC4033] Arends, R., "DNS Security Introduction and Requirements", RFC 4033, March 2005.
- [RFC4655] Farrel, A., Vasseur, J., and J. Ash, "A Path Computation Element (PCE)-Based Architecture", RFC 4655, August 2006.
- [RFC4674] Droms, R., "Requirements for Path Computation Element (PCE) Discovery", RFC 4674, December 2003.
- [RFC5440] Le Roux, J.L., "Path Computation Element (PCE) Communication Protocol (PCEP)", RFC 5440, April 2007.
- [RFC6733] Fajardo, V., "Diameter Base Protocol", RFC 6733, October 2012.
- [RFC6781] Kolkman, O., Mekking, W., and R. Gieben, "DNSSEC Operational Practices, Version 2", RFC 6781,

December 2012.

- [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.

11.2. Informative References

- [ALTO] Kiesel, S., "ALTO Server Discovery", ID draft-ietf-alto-server-discovery-08, March 2013.
- [BGP-LS] Gredler, H., "North-Bound Distribution of Link-State and TE Information using BGP", ID draft-ietf-idr-ls-distribution-03, May 2013.
- [PCE-QUESTION] Farrel, A., "Unanswered Questions in the Path Computation Element Architecture", ID <http://tools.ietf.org/html/draft-ietf-pce-questions-00>, July 2013.
- [RFC2385] Heffernan, A., "Protection of BGP Sessions via the TCP MD5 Signature Option", RFC 2385, August 1998.
- [RFC4927] Le Roux, JL., "Path Computation Element Communication Protocol (PCECP) Specific Requirements for Inter-Area MPLS and GMPLS Traffic Engineering", RFC 4927, June 2007.
- [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.
- [RFC5245] Rosenberg, J., "Interactive Connectivity Establishment (ICE): A Protocol for Network Address Translator (NAT) Traversal for Offer/Answer Protocols", RFC 5245, April 2010.
- [RFC5295] Touch, J., "The TCP Authentication Option", RFC 5295, June 2010.
- [RFC5376] Bitar, N., "Inter-AS Requirements for the Path Computation Element Communication Protocol (PCECP)", RFC 5376, November 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.
- [RFC6457] Takeda, T., "PCC-PCE Communication and PCE Discovery Requirements for Inter-Layer Traffic Engineering", RFC 6457, June 2007.
- [RFC7025] Otani, T., "Requirements for GMPLS Applications of PCE", RFC 7025, September 2013.

Authors' Addresses

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

Email: sunseawq@huawei.com

Dhruv Dhody
Huawei
Leela Palace
Bangalore, Karnataka 560008
INDIA

Email: dhruv.dhody@huawei.com

Daniel King
Old Dog Consulting
UK

Email: daniel@olddog.co.uk

Diego R. Lopez
Telefonica I+D

Email: diego@tid.es

PCE Working Group
Internet-Draft
Intended status: Standards Track
Expires: April 18, 2014

Q. Wu
D. Dhody
Huawei
S. Previdi
Cisco Systems, Inc
October 15, 2013

Extensions to Path Computation Element Communication Protocol (PCEP) for
handling Link Bandwidth Utilization
draft-wu-pce-pcep-link-bw-utilization-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.

Link bandwidth utilization considering the total bandwidth of a link in current use for the forwarding is an important factor to consider during path computation. This document describes extensions to PCEP to consider them as new constraints during 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 April 18, 2014.

Copyright Notice

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

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of

publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Table of Contents

1. Introduction	3
1.1. Requirements Language	3
2. Terminology	3
3. Link Bandwidth Utilization (LBU)	4
4. Link Reserved Bandwidth Utilization (LRBU)	4
5. PCEP Requirements	4
6. PCEP Extensions	5
6.1. BU Object	5
6.1.1. Elements of Procedure	6
6.2. New Objective Functions	6
6.3. PCEP Message	7
7. Other Considerations	9
7.1. Reoptimization Consideration	9
7.2. Inter-domain Consideration	9
7.3. P2MP Consideration	9
7.4. Stateful PCE	10
8. IANA Considerations	10
9. Security Considerations	10
10. Security Considerations	10
11. Manageability Considerations	10
11.1. Control of Function and Policy	10
11.2. Information and Data Models	10
11.3. Liveness Detection and Monitoring	10
11.4. Verify Correct Operations	10
11.5. Requirements On Other Protocols	10
11.6. Impact On Network Operations	11
12. Acknowledgments	11
13. References	11
13.1. Normative References	11
13.2. Informative References	11
Appendix A. Contributor Addresses	12

1. Introduction

Real time link bandwidth utilization is becoming critical in the path computation in some networks. It is important that link bandwidth utilization is factored in during path computation. PCC can request a PCE to provide a path such that it selects under-utilized links. This document extends PCEP [RFC5440] for this purpose.

Traffic Engineering Database (TED) as populated by Interior Gateway Protocol (IGP) contains Maximum bandwidth, Maximum reservable bandwidth and Unreserved bandwidth ([RFC3630] and [RFC3784]). [OSPF-TE-EXPRESS] and [ISIS-TE-EXPRESS] further populate Residual bandwidth and Available bandwidth. Further [ISIS-TE-EXPRESS] also define Bandwidth Utilization.

[Editors Note: [OSPF-TE-EXPRESS] should also be extended in future version for real time link bandwidth utilization]

The links in the path MAY be monitored for changes in the link bandwidth utilization, re-optimization of such path MAY be further requested.

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).

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

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

PCEP: Path Computation Element Protocol.

RSVP: Resource Reservation Protocol

TE LSP: Traffic Engineering Label Switched Path.

3. Link Bandwidth Utilization (LBU)

The bandwidth utilization on a link, forwarding adjacency, or bundled link is populated in the TED (Bandwidth Utilization in [ISIS-TE-EXPRESS]). For a link or forwarding adjacency, bandwidth utilization represent the actual utilization of the link (i.e.: as measured in the router). For a bundled link, bandwidth utilization is defined to be the sum of the component link bandwidth utilization. This includes traffic for both RSVP and non-RSVP.

LBU Percentage is described as the $(LBU / \text{Maximum bandwidth}) * 100$.

4. Link Reserved Bandwidth Utilization (LRBU)

The reserved bandwidth utilization on a link, forwarding adjacency, or bundled link can be calculated from the TED. This includes traffic for only RSVP-TE LSPs.

LRBU can be calculated by using the Residual bandwidth, available bandwidth and LBU. The actual bandwidth by non-RSVP TE traffic can be calculated by subtracting Available Bandwidth from Residual Bandwidth. Once we have the actual bandwidth for non-RSVP TE traffic, subtracting this from LBU would result in LRBU.

LRBU Percentage is described as the $(LRBU / (\text{current reserved bandwidth})) * 100$; where the current reserved bandwidth can be calculated by subtracting Residual bandwidth from Maximum bandwidth.

5. PCEP Requirements

Following requirements associated with bandwidth utilization are identified for PCEP:

1. PCE supporting this draft MUST have the capability to compute end-to-end path with bandwidth utilization constraints. It MUST also support the combination of bandwidth utilization constraint with existing constraints (cost, hop-limit...).
2. PCC MUST be able to request for bandwidth utilization constraint in PCReq message as the boundary condition that should not be crossed for each link in the path.
3. PCC MUST be able to request for bandwidth utilization constraint in PCReq message as an Objective function (OF) [RFC5541] to be

optimized.

4. PCEs are not required to support bandwidth utilization constraint. Therefore, it MUST be possible for a PCE to reject a PCReq message with a reason code that indicates no support for bandwidth utilization constraint.
5. PCEP SHOULD provide mechanism to handle bandwidth utilization constraint in multi-domain (e.g., Inter-AS, Inter-Area or Multi-Layer) environment.
6. PCEP Extensions

This section defines extensions to PCEP [RFC5440] for requirements outlined in Section 5. The proposed solution is used to consider bandwidth utilization during path computation.

6.1. BU Object

The BU (Bandwidth Utilization) 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 TBD.

Two Object-Type values are defined for the BU object:

- o Link Bandwidth Utilization (LBU): BU Object-Type is 1.
- o Link Reserved Bandwidth Utilization (LRBU): BU Object-Type is 2.

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	0	1
Bandwidth Utilization																																									

BU Object Body Format

Bandwidth utilization (32 bits): Represents the bandwidth utilization quantified as a percentage (as described in Section 3 and Section 4). The basic unit is 0.000000023%, with the maximum value 4,294,967,295 representing 98.784247785% ($4,294,967,295 * 0.000000023\%$). This value is the maximum Bandwidth utilization

percentage that can be expressed.

The BU object body has a fixed length of 4 bytes.

6.1.1. Elements of Procedure

PCC SHOULD request the PCE to factor in the bandwidth utilization during path computation by including a BU object in the PCReq message.

Multiple BU objects MAY be inserted in a PCReq or a PCRep message for a given request but there MUST be at most one instance of the BU object for each object type. If, for a given request, two or more instances of a BU object with the same object type are present, only the first instance MUST be considered and other instances MUST be ignored.

BU object MAY be carried in a PCRep message in case of unsuccessful path computation along with a NO-PATH object to indicate the constraints that could not be satisfied.

If the P bit is clear in the object header and PCE does not understand or does not support bandwidth utilization during path computation it SHOULD simply ignore BU object.

If the P Bit is set in the object header and PCE receives BU object in path request and it understands the BU object, but the PCE is not capable of bandwidth utilization check during path computation, the PCE MUST send a PCErr message with a PCEP-ERROR Object Error-Type = 4 (Not supported object) [RFC5440]. The path computation request MUST then be cancelled.

If the PCE does not understand the BU object, then the PCE MUST send a PCErr message with a PCEP-ERROR Object Error-Type = 3 (Unknown object) [RFC5440].

6.2. New Objective Functions

This document defines two additional objective functions -- namely, MUP (Maximum Under-Utilized Path) and MRUP (Maximum Reserved Under-Utilized Path). Hence two new objective function codes have to be defined.

Objective functions are formulated using the following terminology:

- o A network comprises a set of N links $\{L_i, (i=1...N)\}$.

- o A path P is a list of K links $\{L_{pi}, (i=1...K)\}$.
- o Bandwidth Utilization on link L is denoted $u(L)$.
- o Reserved Bandwidth Utilization on link L is denoted $ru(L)$.
- o Maximum bandwidth on link L is denoted $M(L)$.
- o Current Reserved bandwidth on link L is denoted $c(L)$.

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

Objective Function Code: TBD

Name: Maximum Under-Utilized Path (MUP)

Description: Find a path P such that $(\text{Min } \{(M(L_{pi}) - u(L_{pi})) / M(L_{pi}), i=1...K\})$ is maximized.

Objective Function Code: TBD

Name: Maximum Reserved Under-Utilized Path (MRUP)

Description: Find a path P such that $(\text{Min } \{(c(L_{pi}) - ru(L_{pi})) / c(L_{pi}), i=1...K\})$ is maximized.

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

If the objective function defined in this document are unknown/unsupported, the procedure as defined in [RFC5541] is followed.

6.3. PCEP Message

The new optional BU objects MAY be specified in the PCReq message. As per [RFC5541], an OF object specifying a new objective function MAY also be specified.

The format of the PCReq message (with [RFC5541] 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>
                [<LSPA>]
                [<BANDWIDTH>]
                [<bu-list>]
                [<metric-list>]
                [<OF>]
                [<RRO>[<BANDWIDTH>]]
                [<IRO>]
                [<LOAD-BALANCING>]

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

```

The BU objects MAY be specified in the PCRep message, in case of an unsuccessful path computation to indicate the bandwidth utilization as a reason for failure. The OF object MAY be carried within a PCRep message to indicate the objective function used by the PCE during path computation.

The format of the PCRep message (with [RFC5541] 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>
                [<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>]

```

7. Other Considerations

7.1. Reoptimization Consideration

PCC can monitor the link bandwidth utilization of the setup LSPs and in case of drastic change, it MAY ask PCE for reoptimization as per [RFC5440].

7.2. Inter-domain Consideration

7.3. P2MP Consideration

7.4. Stateful PCE

8. IANA Considerations

TBD

9. Security Considerations

TBD

10. Security Considerations

This document defines a new BU object and OF codes which does not add any new security concerns beyond those discussed in [RFC5440].

11. Manageability Considerations

11.1. 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. If the new constraints are not supported/allowed on a PCE, it MUST send a PCErr message as specified in Section 6.1.1.

11.2. Information and Data Models

[PCEP-MIB] describes the PCEP MIB, there are no new MIB Objects for this document.

11.3. Liveness Detection and Monitoring

Mechanisms defined in this document do not imply any new liveness detection and monitoring requirements in addition to those already listed in [RFC5440].

11.4. Verify Correct Operations

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

11.5. Requirements On Other Protocols

PCE requires the TED to be populated with the bandwidth utilization. This mechanism is described in [OSPF-TE-EXPRESS] or [ISIS-TE-EXPRESS].

11.6. Impact On Network Operations

Mechanisms defined in this document do not have any impact on network operations in addition to those already listed in [RFC5440].

12. Acknowledgments

We would like to thank Alia Atlas, John E Drake, David Ward for their useful comments and suggestions.

13. References

13.1. Normative References

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

13.2. Informative References

- [RFC3630] Katz, D., Kompella, K., and D. Yeung, "Traffic Engineering (TE) Extensions to OSPF Version 2", RFC 3630, September 2003.
- [RFC3784] Smit, H. and T. Li, "Intermediate System to Intermediate System (IS-IS) Extensions for Traffic Engineering (TE)", RFC 3784, June 2004.
- [RFC5440] Vasseur, JP. and JL. Le Roux, "Path Computation Element (PCE) Communication Protocol (PCEP)", RFC 5440, March 2009.
- [RFC5541] Le Roux, JL., Vasseur, JP., and Y. Lee, "Encoding of Objective Functions in the Path Computation Element Communication Protocol (PCEP)", RFC 5541, June 2009.
- [OSPF-TE-EXPRESS] Giacalone, S., Ward, D., Drake, J., Atlas, A., and S. Previdi, "OSPF Traffic Engineering (TE) Metric Extensions [draft-ietf-ospf-te-metric-extensions]", June 2013.
- [ISIS-TE-EXPRESS] Previdi, S., Giacalone, S., Ward, D., Drake, J., Atlas, A., Filsfils, C., and W. Qin, "IS-IS Traffic Engineering (TE) Metric Extensions [draft-ietf-isis-te-metric-extensions-01]", October 2013.

[PCEP-MIB] Kiran Koushik, A S., Stephan, E., Zhao, Q., King,
D., and J. Hardwick, "PCE communication
protocol(PCEP) Management Information Base
[draft-ietf-pce-pcep-mib]", Feb 2013.

Appendix A. Contributor Addresses

Udayasree Palle
Huawei Technologies
Leela Palace
Bangalore, Karnataka 560008
INDIA
EMail: udayasree.palle@huawei.com

Authors' Addresses

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

EMail: sunseawq@huawei.com

Dhruv Dhody
Huawei Technologies
Leela Palace
Bangalore, Karnataka 560008
INDIA

EMail: dhruv.ietf@gmail.com

Stefano Previdi
Cisco Systems, Inc
Via Del Serafico 200
Rome 00191
IT

EMail: sprevidi@cisco.com

Network Working Group
Internet-Draft
Intended status: Standards Track

Xian Zhang
Young Lee
Fatai Zhang
Huawei
Ramon Casellas
CTTC
Oscar Gonzalez de Dios
Telefonica I+D
Zafar Ali
Cisco Systems

Expires: April 21, 2014

October 21, 2013

Path Computation Element (PCE) Protocol Extensions for Stateful PCE
Usage in GMPLS-controlled Networks

draft-zhang-pce-pcep-stateful-pce-gmpls-03.txt

Status of this Memo

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

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

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

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

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

This Internet-Draft will expire on April 21, 2014.

Abstract

The Path Computation Element (PCE) facilitates Traffic Engineering (TE) based path calculation in large, multi-domain, multi-region, or multi-layer networks. [Stateful-PCE] provides the fundamental PCE communication Protocol (PCEP) extensions needed to support stateful PCE functions, without specifying the technology-specific extensions. This memo provides extensions required for PCEP so as to enable the usage of a stateful PCE capability in GMPLS-controlled networks.

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

Table of Contents	2
1. Introduction	3
2. PCEP Extensions	3
2.1. Overview of Requirements.....	3
2.2. Stateful PCE Capability Advertisement	4
2.2.1. PCE Capability Advertisement in Multi-layer Networks	4
2.3. LSP Delegation in GMPLS-controlled Networks	5
2.4. LSP Synchronization in GMPLS-controlled networks.....	6
2.5. Modification of Existing PCEP Messages and Procedures....	7
2.5.1. Use cases	8
2.5.2. Modification for LSP Re-optimization	8
2.5.3. Modification for Route Exclusion	9
2.6. Additional Error Type and Error Values Defined.....	10
3. IANA Considerations	10
4. Manageability Considerations	10
4.1. Requirements on Other Protocols and Functional Components	10
5. Security Considerations.....	11
6. Acknowledgement	11
7. References	11
7.1. Normative References.....	11
7.2. Informative References.....	11
8. Contributors' Address.....	12
Authors' Addresses	13

1. Introduction

[RFC 4655] presents the architecture of a Path Computation Element (PCE)-based model for computing 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). To request path computation services to a PCE, [RFC 5440] defines the PCE communication Protocol (PCEP) for interaction between a Path Computation Client (PCC) and a PCE, or between two PCEs. PCEP as specified in [RFC 5440] mainly focuses on MPLS networks and the PCEP extensions needed for GMPLS-controlled networks are provided in [PCEP-GMPLS].

Stateful PCEs are shown to be helpful in many application scenarios, in both MPLS and GMPLS networks, as illustrated in [Stateful-APP]. In order for these applications to be able to exploit the capability of stateful PCEs, extensions to the PCE communication protocol (i.e., PCEP) are required.

[Stateful-PCE] provides the fundamental extensions needed for stateful PCE to support general functionality, but leaves out the specification for technology-specific objects/TLVs. Complementarily, this document focuses on the extensions that are necessary in order for the deployment of stateful PCEs in GMPLS-controlled networks.

2. PCEP Extensions

2.1. Overview of Requirements

This section notes the main functional requirements for PCEP extensions to support stateful PCE for use in GMPLS-controlled networks, based on the description in [Stateful-APP]. Many requirements are common across a variety of network types (e.g., MPLS-TE networks and GMPLS networks) and the protocol extensions to meet the requirements are already described in [Stateful-PCE]. This document does not repeat the description of those protocol extensions. Other requirements that are also common across a variety of network types do not currently have protocol extensions defined in [Stateful-PCE]. In these cases, this document presents protocol extensions for discussion by the PCE working group and potential inclusion in [Stateful-PCE]. In addition, this document presents protocol extensions for a set of requirements which are specific to the use of a stateful PCE in a GMPLS-controlled network.

The basic requirements are as follows:

- o Advertisement of the stateful PCE capability. This generic requirement is covered in Section 7.1.1 of [Stateful-PCE]. Section 2.2 of this document discusses other potential extensions for this functionality.
- o LSP delegation is already covered in Section 5.5 of [Stateful-PCE]. Section 2.3 of this document provides extension for its application in GMPLS-controlled networks. Moreover, further discussion of some generic details that may need additional consideration is provided.
- o LSP state synchronization. This is a generic requirement already covered in Section 5.4 of [Stateful-PCE]. However, there are further extensions required specifically for GMPLS-controlled networks and discussed in Section 2.4. Reference to LSPs by identifiers is discussed in Section 7.2 of [Stateful-PCE]. This feature can be applied to reduce the data carried in PCEP messages. Use cases and additional Error Codes are necessary, as described in Section 2.5 and 2.6.

2.2. Stateful PCE Capability Advertisement

Whether a PCE has stateful capability or not can be advertised during the PCEP session establishment process. It can also be advertised through routing protocols as described in [RFC5088]. In either case, the following additional aspects should also be considered.

2.2.1. PCE Capability Advertisement in Multi-layer Networks

In multi-layer network scenarios, such as an IP-over-optical network, if there are dedicated PCEs responsible for each layer, then the PCCs should be informed of which PCEs they should synchronize their LSP states with, as well as send path computation requests to. The Layer-Cap TLV defined in [INTER-LAYER] can be used to indicate which layer a PCE is in charge of. (Editor's note: this change is currently not included in the current version of the [INTER-LAYER] draft. It is expected that it will be included in its next version.) This TLV is optional and MAY be carried in the OPEN object. It is RECOMMENDED that a PCC synchronizes its LSP states with the same PCEs that it can use for path computation in a multi-layer network. In a single layer, this TLV MAY not be used. However, if the PCE capability discovery depends on IGP and if an IGP instance spans across multiple layers, this TLV is still needed.

Alternatively, the extension to current OSPF PCED TLV is needed. A new domain-type denoting the layer information can be defined:

domain-type: T.B.D.

When it is carried in PCE-DOMAIN sub-TLV, it denotes the layer for which a PCE is responsible for path computation as well as LSP state synchronization. When carried in the PCE-NEIG-DOMAIN sub-TLV, it denotes its adjacent layers for which a PCE can compute paths and synchronize the LSP states. The DOMAIN-ID information can be represented using the following format, to denote the layer information:

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								
LSP Enc. Type										Switching Type										Reserved																			

2.3. LSP Delegation in GMPLS-controlled Networks

To enable the PCE to control an LSP, the PCUpd message is defined in [Stateful-PCE]. However, the specification of technology specific extensions is not covered. The following defines the <path> descriptor, present in the PCUpd message, that should be used in GMPLS-controlled networks:

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

Where:

```

<attribute-list> ::= [ <LSPA> ]
                    [ <BANDWIDTH> ]
                    [ <GENERALIZED-BANDWIDTH>... ]
                    [ <metric-list> ]

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

```

As explained in [stateful-APP], LSP parameter update controlled by a stateful PCE in a multi-domain network is complex and requires well-defined operational procedures as well as protocol design.

[TBD: protocol extensions]

2.4. LSP Synchronization in GMPLS-controlled networks

For LSP state synchronization of stateful PCEs in GMPLS networks, the LSP attributes, such as its bandwidth, associated route as well as protection information etc, should be updated by PCCs to PCE LSP database (LSP-DB). Note the LSP state synchronization described in this document denotes both the bulk LSP report at the initialization phase as well as the LSP state report afterwards described in [Stateful-PCE].

As per [Stateful-PCE], it does not cover technology-specific specification for state synchronization. Therefore, extensions of PCEP for stateful PCE usage in GMPLS networks are required. For LSP state synchronization, the objects/TLVs that should be used for stateful PCE in GMPLS networks are defined in [PCEP-GMPLS] and are briefly summarized as below:

- o GENERALIZED BANDWIDTH
- o GENERALIZED ENDPOINTS
- o PROTECTION ATTRIBUTE
- o Use of IF_ID_ERROR_SPEC. [Stateful-PCE] section 7.2.2 only considers RSVP_ERROR_SPEC TLVs. GMPLS extends this to also support IF_ID_ERROR_SPEC, for example, to report about failed unnumbered interfaces.
- o Extended objects to support the inclusion of the label and unnumbered links.

Per [Stateful-PCE], the PCRpt message is defined for LSP state synchronization purposes. PCRpt is used by a PCC to report one or more of its LSPs to a stateful PCE. However, the <path> descriptor is technology-specific and left undefined.

For LSP state synchronization in GMPLS-controlled networks, the encoding of the <path> descriptor is defined as follows:

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

Where:

```
<attribute-list> ::= [ <LSPA> ]
                        [ <BANDWIDTH> ]
                        [ <GENERALIZED-BANDWIDTH> ... ]
```

[<IRO>]

[<XRO>]

[<metric-list>]

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

The objects included in the <path> descriptor can be found in [RFC5440], [PCE-GMPLS] and [RFC5521].

For all the objects presented in this section, the P and I bit MUST be set to 0 since they are only used by a PCC to report its LSP information.

In GMPLS-controlled networks, the <ERO> object may include a list of the label sub-object for SDH/SONET, OTN and DWDM networks. It may also include a list of unnumbered interface IDs to denote the allocated resource. The <RRO>, <IRO> and <XRO> objects MAY include unnumbered interface IDs and labels for networks such as OTN and WDM networks.

If the LSP being reported is a protecting LSP, the <PROTECTION-ATTRIBUTE> TLV MUST be included in the <LSPA> object to denote its attributes and restrictions. Moreover, if the status of the protecting LSP changes from non-operational to operational, this should be synchronized to the stateful PCE. For example, in 1:1 protection, the combination of S=0, P=1 and O=0 denotes the protecting path is set up already but not used for carrying traffic. Upon the working path failure, the operational status of the aforementioned protecting LSP changes to in-use (i.e., O=1). This information should be synchronized with a stateful PCE through a PCRpt message.

The O bit in the <GENERALIZED-BANDWIDTH> object has no meaning for LSP state synchronization and MUST be set to 0. Furthermore, this object MAY appear twice, one with R set to 1 and the other with R set to 0. This is to denote the asymmetric bandwidth property of the updated bi-directional LSP.

2.5. Modification of Existing PCEP Messages and Procedures

One of the advantages mentioned in [Stateful-APP] is that the stateful nature of a PCE simplifies the information conveyed in PCEP messages, notably between PCC and PCE, since it is possible to refer to PCE managed state for active LSPs. To be more specific, with a

stateful PCE, it is possible to refer to a LSP with a unique identifier in the scope of the PCC-PCEP session and thus use such identifier to refer to that LSP.

2.5.1. Use cases

Use Case 1: Assuming a stateful PCE's LSP-DB is up-to-date, a PCC (e.g. NMS) requesting for a re-optimization of one or several LSPs can send the request with "R" bit set and only provides the relevant LSP unique identifiers.

Upon receiving the PCReq message, PCE should be able to correlate with one or multiple LSPs with their detailed state information and carry out optimization accordingly.

The handling of RP object specified in [RFC5440] is stated as following:

"The absence of an RRO in the PCReq message for a non-zero-bandwidth TE LSP (when the R bit of the RP object is set) MUST trigger the sending of a PCErr message with Error-Type="Required Object Missing" and Error-value="RRO Object missing for re-optimization."

If a PCE has stateful capabilities, and such capabilities have been negotiated and advertised, specific rules given in [RFC5440] may need to be relaxed. In particular, the re-optimization case: if the re-optimization request refers to a given LSP state, and the RRO information is available, the PCE can proceed.

Use Case 2: in order to set up a LSP which has a constraint that its route should not use resources used by one or more existing LSPs, a PCC can send a PCReq with the identifiers of these LSPs. A stateful PCE should be able to find the corresponding route and resource information so as to meet the constraints set by the requesting PCC. Hence, the LSP identifier TLV defined in [Stateful-PCE] can be used in XRO object for this purpose. Note that if the PCC is a node in the network, the constraint LSP ID information will be confined to the LSPs initiated by itself.

2.5.2. Modification for LSP Re-optimization

For re-optimization, upon receiving a path computation request and the "R" bit is set, the stateful PCE SHOULD still perform the re-optimization in the following two cases:

Case 1: the existing bandwidth and route information of the to-be-optimized LSP is provided in the path computation request. This

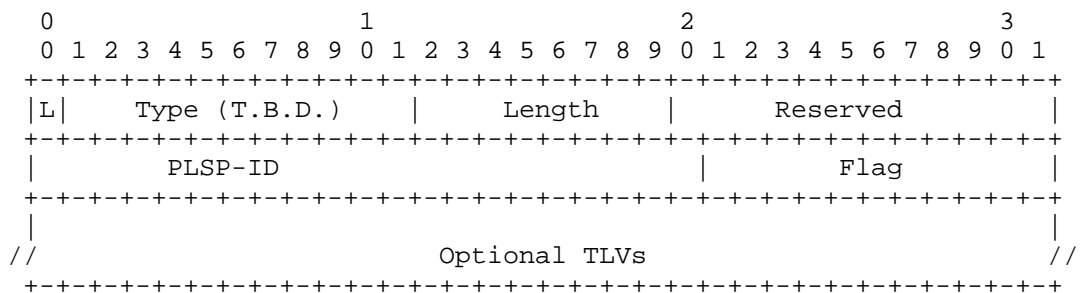
information should be provided via <BANDWIDTH>, <GENERALIZED-BANDWIDTH>, <ERO> objects.

Case 2: the existing bandwidth and route information can be found locally in its LSP-DB. In this case, the PCRep and PCReq messages need to be modified to carry LSP identifiers. The stateful PCE can find this information using the per-node LSP ID together with the PCC's address.

If no LSP state information is available to carry out re-optimization, the stateful PCE should report the error "LSP state information unavailable for the LSP re-optimization" (Error Type = T.B.D., Error value= T.B.D.).

2.5.3. Modification for Route Exclusion

A LSP identifier sub-object is defined and its format as follows:



L bit:

The L bit SHOULD NOT be set, so that the subobject represents a strict hop in the explicit route.

Type:

Subobject Type for a per-node LSP identifier.

Length:

The Length contains the total length of the subobject in bytes, including the Type and Length fields.

PLSP-ID:

This is the identifier given to a LSP and it is unique on a node basis. It is defined in [Stateful-PCE].

Flags:

This field is defined in [Stateful-PCE]. It is not used in this sub-object and should be ignored upon receipt.

Optional TLVs:

Additional TLVs can be defined in the future to provide further information to identify a LSP. In this document, no TLVs are defined.

One or multiple of these sub-objects can be present in the XRO object. When a stateful PCE receives a path computation request carrying this sub-object, it should find relevant information of these LSPs and preclude the resource during the path computation process. If a stateful PCE cannot recognize one or more of the received LSP identifiers, it should reply PCErr saying "the LSP state information for route exclusion purpose cannot be found" (Error-type = T.B.D., Error-value= T.B.D.). Optionally, it may provide with the unrecognized identifier information to the requesting PCC.

2.6. Additional Error Type and Error Values Defined

Error Type Meaning

21(TBD) LSP state information missing

Error-value 1: LSP state information unavailable for the LSP re-optimization

Error-value 2: the LSP state information for route exclusion purpose cannot be found

3. IANA Considerations

IANA is requested to allocate new Types for the TLV/Object defined in this document.T.B.D.

4. Manageability Considerations

The description and functionality specifications presented related to stateful PCEs should also comply with the manageability specifications covered in Section 8 of [RFC4655]. Furthermore, a further list of manageability issues presented in [Stateful-PCE] should also be considered.

Additional considerations are presented in the next sections.

4.1. Requirements on Other Protocols and Functional Components

When the detailed route information is included for LSP state synchronization (either at the initial stage or during LSP state

report process), this require the ingress node of an LSP carry the RRO object in order to enable the collection of such information.

5. Security Considerations

The security issues presented in [RFC5440] and [Stateful-PCE] apply to this document.

6. Acknowledgement

We would like to thank Adrian Farrel and Cyril Margaria for the useful comments and discussions.

7. References

7.1. Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to indicate requirements levels", RFC 2119, March 1997.
- [RFC4655] Farrel, A., Vasseur, J.-P., and Ash, J., "A Path Computation Element (PCE)-Based Architecture", RFC 4655, August 2006.
- [RFC5440] Vasseur, J.-P., and Le Roux, JL., "Path Computation Element (PCE) Communication Protocol (PCEP)", RFC 5440, March 2009.
- [RFC5088] Le Roux, JL., Vasseur, J.-P., Ikejiri, Y., Zhang, R., "OSPF Protocol Extensions for Path Computation Element (PCE) Discovery", RFC 5088, January 2008.
- [INTER-LAYER] Oki, E., Takeda, Tomonori, Le Roux, JL., Farrel, A., Zhang, F., "Extensions to the Path Computation Element communication Protocol (PCEP) for Inter-Layer MPLS and GMPLS Traffic Engineering", draft-ietf-pce-inter-layer-ext, work in progress.

7.2. Informative References

- [Stateful-APP] Zhang, X., Minei, I., et al "Applicability of Stateful Path Computation Element (PCE) ", draft-ietf-pce-stateful-pce-app, , work in progress.
- [Stateful-PCE] Crabbe, E., Medved, J., Varga, R., Minei, I., "PCEP Extensions for Stateful PCE", draft-ietf-pce-stateful-pce, work in progress.

[PCE-IA-WSN] Lee, Y., Bernstein G., Takeda, T., Tsuritani, T.,
"PCEP Extensions for WSON Impairments", draft-lee-pce-
wson-impairments, work in progress.

[PCEP-GMPLS] Margaria, C., Gonzalez de Dios, O., Zhang, F., "PCEP
extensions for GMPLS", draft-ietf-pce-gmpls-pcep-
extensions, work in progress.

8. Contributors' Address

Dhruv Dhody
Huawei Technology
Leela Palace
Bangalore, Karnataka 560008
INDIA

EMail: dhruvd@huawei.com

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

Phone: +86-755-28972914
Email: yi.lin@huawei.com

Authors' Addresses

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

Phone: +86-755-28972645
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: ylee@huawei.com

Fatai Zhang
Huawei
F3-5-B R&D Center, Huawei Base
Bantian, Longgang District
P.R. China

Phone: +86-755-28972912
Email: zhangfatai@huawei.com

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

Phone:
Email: ramon.casellas@cttc.es

Oscar Gonzalez de Dios
Telefonica Investigacion y Desarrollo
Emilio Vargas 6
Madrid, 28045
Spain

Phone: +34 913374013
Email: ogondio@tid.es

Zafar Ali
Cisco Systems
Email: zali@cisco.com

Intellectual Property

The IETF Trust takes no position regarding the validity or scope of any Intellectual Property Rights or other rights that might be claimed to pertain to the implementation or use of the technology described in any IETF Document or the extent to which any license under such rights might or might not be available; nor does it represent that it has made any independent effort to identify any such rights.

Copies of Intellectual Property disclosures made to the IETF Secretariat and any assurances of licenses to be made available, or the result of an attempt made to obtain a general license or permission for the use of such proprietary rights by implementers or users of this specification can be obtained from the IETF on-line IPR repository at <http://www.ietf.org/ipr>

The IETF invites any interested party to bring to its attention any copyrights, patents or patent applications, or other proprietary rights that may cover technology that may be required to implement any standard or specification contained in an IETF Document. Please address the information to the IETF at ietf-ipr@ietf.org.

The definitive version of an IETF Document is that published by, or under the auspices of, the IETF. Versions of IETF Documents that are published by third parties, including those that are translated into other languages, should not be considered to be definitive versions of IETF Documents. The definitive version of these Legal Provisions is that published by, or under the auspices of, the IETF. Versions of these Legal Provisions that are published by third parties, including those that are translated into other languages, should not be considered to be definitive versions of these Legal Provisions.

For the avoidance of doubt, each Contributor to the IETF Standards Process licenses each Contribution that he or she makes as part of the IETF Standards Process to the IETF Trust pursuant to the provisions of RFC 5378. No language to the contrary, or terms,

conditions or rights that differ from or are inconsistent with the rights and licenses granted under RFC 5378, shall have any effect and shall be null and void, whether published or posted by such Contributor, or included with or in such Contribution.

Disclaimer of Validity

All IETF Documents and the information contained therein are provided on an "AS IS" basis and THE CONTRIBUTOR, THE ORGANIZATION HE/SHE REPRESENTS OR IS SPONSORED BY (IF ANY), THE INTERNET SOCIETY, THE IETF TRUST AND THE INTERNET ENGINEERING TASK FORCE DISCLAIM ALL WARRANTIES, EXPRESS OR IMPLIED, INCLUDING BUT NOT LIMITED TO ANY WARRANTY THAT THE USE OF THE INFORMATION THEREIN WILL NOT INFRINGE ANY RIGHTS OR ANY IMPLIED WARRANTIES OF MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE.

Full Copyright Statement

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.

PCE Working Group
Internet-Draft
Intended status: Standards Track
Expires: April 24, 2014

Quintin Zhao
Katherine Zhao
Robin Li
Huawei Technologies
Zekun Ke
Tencent Holdings Ltd.
October 21, 2013

The User cases for Using PCE as the Central Controller(PCECC) of LSPs
draft-zhao-pce-central-controller-user-cases-00

Abstract

In certain deployment networks deployment scenarios, service providers would like to keep all the existing MPLS functionalities in both MPLS and GMPLS network while removing the complexity of existing signaling protocols such as LDP and RSVP-TE. In this document, we propose to use the PCE as a central controller so that LSP can be calculated/signaled/initiated/downloaded through a centralized PCE server to each network devices along the LSP path while leveraging the existing PCE technologies as much as possible.

This draft describes the user cases for using the PCE as the central controller where LSPs are calculated/setup/initiated/downloaded through extending the existing PCE architectures and extending the 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 April 24, 2014.

Copyright Notice

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

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

Table of Contents

1. Introduction	3
1.1. Using the PCE as the Central Controller (PCECC) Approach	4
1.2. MPLS Label Resource Reservation through PCECCP	5
1.3. Using PCECCP to Distribute the LSP Forwarding Entry from PCECC server to each PCECC clients	6
2. Terminology	6
3. PCEP Requirements	6
4. User Cases for PCECC's Label Resource Reservations	7
5. User Cases for PCECC for LSP Setup in the New PCECC Enabled Network	8
6. User Cases for PCECC for LSP Setup in the Network Migration .	8
7. Using Extended PCEP to download LSP infor for Each Network Device	9
8. The Considerations for PCECC Procedure and PCEP extensions . .	10
9. Acknowledgments	10
10. References	10
10.1. Normative References	10
10.2. Informative References	11

1. Introduction

In certain network deployment scenarios, service providers would like to have the ability to dynamically adapt to a wide range of customer's requests for the sake of flexible network service delivery, SDN has provides additional flexibility in how the network is operated comparing the traditional network.

The existing networking ecosystem has become awfully complex and highly demanding in terms of robustness, performance, scalability, flexibility, agility, etc. By migrating to the SDN enabled network from the existing network, service providers and network operators must have a solution which they can evolve easily from the existing network into the SDN enabled network while keeping the network services remain scalable, guarantee robustness and availability etc.

Taking the smooth transition between traditional network and the new SDN enabled network into account, especially from a cost impact assessment perspective, using the existing PCE components from the current network to function as the central controller of the SDN network is one choice, which not only achieves the goal of having a centralized controller to provide the functionalities needed for the central controller, but also leverages the existing PCE network components.

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 MPLS network that is centrally controlled and deployed.

[I-D.draft-ali-pce-remote-initiated-gmpls-lsp-01] complements [I-D. draft-crabbe-pce-pce-initiated-lsp] by addressing the requirements for remote-initiated GMPLS LSPs.

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.

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.

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 a PCC using PCEP extensions specified in [I-D.crabbe-pce-pce-initiated-lsp] using the SR specific PCEP extensions described in [I-D.draft-sivabalan-pce-segment-routing].

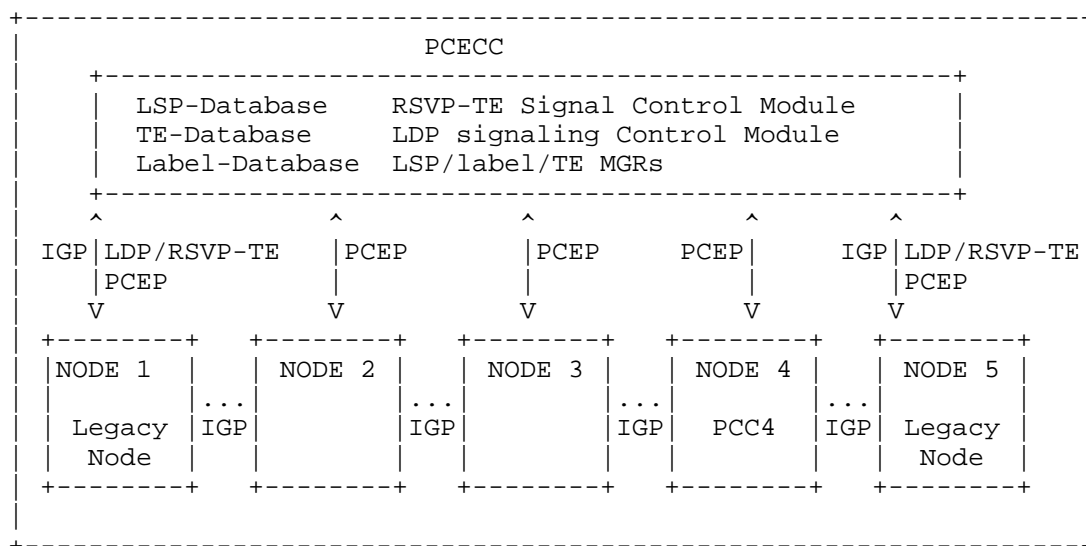
By using the solutions provided from above drafts, LSP in both MPLS and GMPLS network can be setup/delete/maintained/synchronized through a centrally controlled dynamic MPLS network. Since in these solutions, the LSP is need to be signaled through the head end LER to the tail end LER, there are either RSVP-TE signaling protocol need to be deployed in the MPLS/GMPLS network, or extend TGP protocol with node/adjacency segment identifiers signaling capability to be deployed.

The PCECC solution proposed in this document allow for a dynamic MPLS network that is eventually controlled and deployed without the deployment of RSVP-TE protocol or extended IGP protocol with node/adjacency segment identifiers signaling capability while providing all the key MPLS functionalities needed by the service providers. In the case that one LSP path consists legacy network nodes and the new network nodes which are centrally controlled, the PCECC solution provides a smooth transition step for users.

1.1.1. Using the PCE as the Central Controller (PCECC) Approach

With PCECC, it not only removes the existing MPLS signaling totally from the control plane without losing any existing MPLS functionalities, but also PCECC achieves this goal through utilizing the existing PCEP without introducing a new protocol into the network.

The following diagram illustrates the PCECC architecture.



Through the draft, we call the combination of the functionality for global label range signaling and the functionality of LSP setup/download/cleanup using the combination of global labels and local labels as PCECC functionality.

1.2. MPLS Label Resource Reservation through PCECCP

Current MPLS label has local meaning. That is, MPLS label allocated locally and signaled through the LDP/RSVP-TE/BGP etc dynamic signaling protocol.

As the SDN(Service-Driven Network) technology develops, MPLS global label has been proposed again for new solutions. [I-D.li-mppls-global-label-usecases] proposes possible usecases of MPLS global label. MPLS global label can be used for identification of the location, the service and the network in different application scenarios. From these usecases we can see that no matter SDN or traditional application scenarios, the new solutions based on MPLS global label can gain advantage over the existing solutions to facilitate service provisions.

To ease the label allocation and signaling mechanism, also with the new applications such as concentrated LSP controller is introduced, PCE can be conveniently used as a central controller and MPLS global label range negotiator.

The later section of this draft describes the user cases for PCE

server and PCE clients to have the global label range negotiation and local label range negotiation functionality.

1.3. Using PCECCP to Distribute the LSP Forwarding Entry from PCECC server to each PCECC clients

To empower networking with centralized controllable modules, there are many choices for downloading the forwarding entries to the data plane, one way is the use of the OpenFlow protocol, which helps devices populate their forwarding tables according to a set of instructions to the data plane. There are other candidate protocols to convey specific configuration information towards devices also. Since the PCEP protocol is already deployed in some of the service network, to leverage the PCEP to populated the MPLS forwarding table is a possible good choice.

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).

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

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

TE: Traffic Engineering.

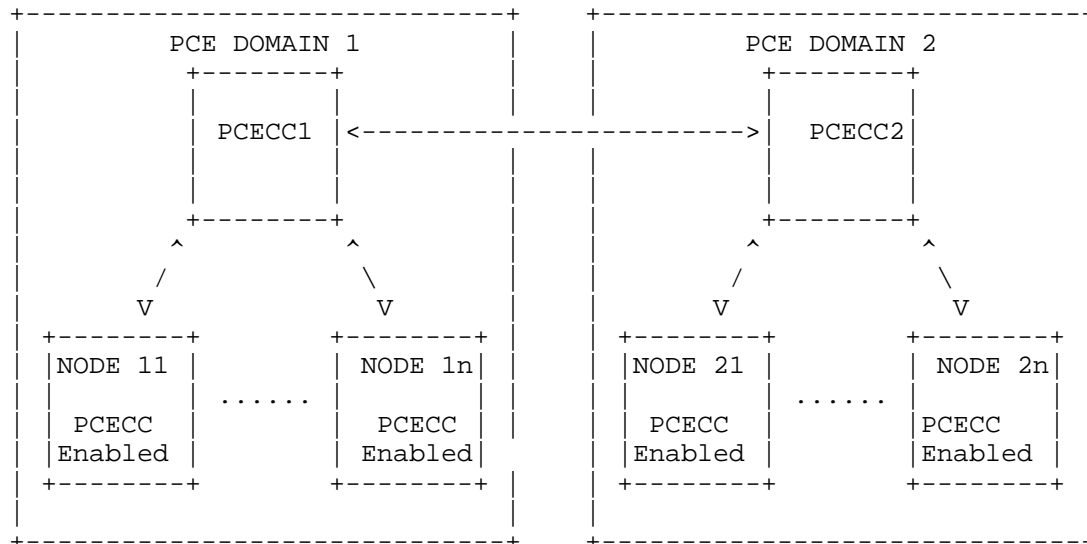
3. PCEP Requirements

Following key requirements associated PCECC should be considered when designing the PCECC based solution:

1. Path Computation Element (PCE) clients supporting this draft MUST have the capability to advertise its PCECC capability to the PCECC.
2. Path Computation Element (PCE) supporting this draft MUST have the capability to negotiate a global label range for a group of clients.

3. Path Computation Client (PCC) MUST be able ask for global label range assigned in path request message .
 4. PCE are not required to support label reserve service. Therefore, it MUST be possible for a PCE to reject a Path Computation Request message with a reason code that indicates no support for label reserve service.
 5. PCEP SHOULD provide a means to return global label range and LSP label assignments of the computed path in the reply message.
 6. PCEP SHOULD provide a means to download the MPLS forwarding entry to the PCECC's clients.
4. User Cases for PCECC's Label Resource Reservations

Example 1 to 3 are based on network configurations illustrated using the following figure:



Example 1: global Label Range Reservation

- o Node11 sends a label request message to PCECC1 with global label reservation range 100 to 1000.
- o PCECC1 sends a reply message with global label reservation range 100 to 1000 confirmed to node1, ..., node1n

- o PCECC1 sends a indication message with global label reservation range 100 to 1000 confirmed to PCECC2.
- o PCECC2 sends indication messages with global label reservation range 100 to 1000 confirmed to Node21,..., node2n

5. User Cases for PCECC for LSP Setup in the New PCECC Enabled Network

Example 2: Tunnel Head End Initiated LSP Setup Using Global Label Range Reserved

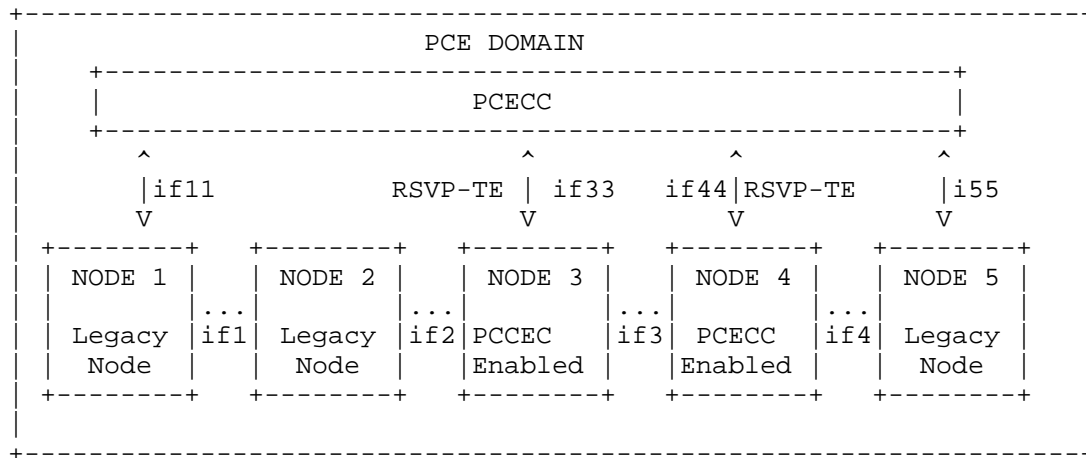
- o Node1 sends a path request message for LSP setup from Node11 to Node2n.
- o PCE1 sends a indication message LSP setup with [Label(to2n), Node2n] to Node12, ..., Node1n.
- o PCE1 sends a indication message LSP setup with [Label(to2n), Node2n] to PCE2;
- o PCE2 sends a indication message LSP setup with [Label(to2n), Node2n] to Node22, ..., Node2n.

Example 3: LSP Delete Using global Label Range Reserved

- o Node1 sends a path request message for LSP cleanup from Node11 to Node2n.
- o PCE1 sends a indication message LSP cleanup with [Label(to2n), Node2n] to Node12, ..., Node1n.
- o PCE1 sends a indication message LSP cleanup with [Label(to2n), Node2n] to PCE2;
- o PCE2 sends a indication message LSP cleanup with [Label(to2n), Node2n] to Node22, ..., Node2n.

6. User Cases for PCECC for LSP Setup in the Network Migration

Example 4 is based on network configurations illustrated using the following figure:



Example 4: PCECC Initiated LSP Setup In the Network Migration

In this example, there five nodes for the LSP from head end (node1) to the tail end (node5). Where the node3 and node4 with the PCECC capability, and other nodes are legacy nodes.

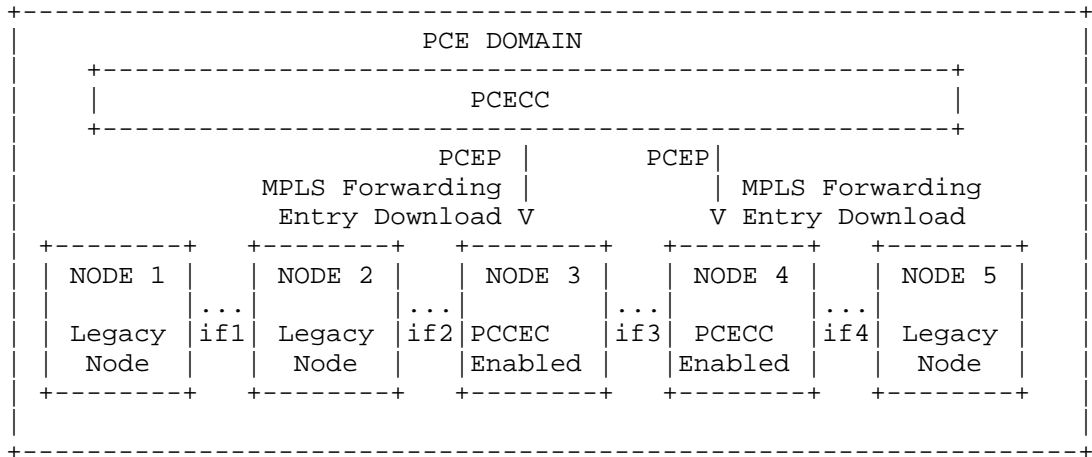
- o Node1 sends a path request message for LSP setup to Node5.
- o PCE sends a reply message for LSP setup with path (node1, i1), (node2, i2), (node-PCECC, if33), (node-PCECC, if55), Nnode5.
- o PCE sends an indication message for LSP segment setup with [Label(toN5), Node5] for node3 to node4.
- o Node1, Node2, Node-PCECC, Node-PCECC, Node 5 will setup the LSP to Node5 normally using the local label as normal. After the LSP is setup, then the PCECC will program the node 3 and node 4 to replace the LSP segment from node3-node-pcecc-node5 to node3-node4-node5.

7. Using Extended PCEP to download LSP info for Each Network Device

The existing PCEP is used to communicate between the PCE server and PCE's client PCC for exchanging the path request and reply information regarding to the LSP info. With minor extensions, we can use the PCEP to download the complete LSP forwarding entries for each node in the network.

In the example 4, the LSP segment between node3 and node4 for destination node5 is setup from PCECC and downloaded into node3 and

node4 directly from PCECC through the extended PCEP.



8. The Considerations for PCECC Procedure and PCEP extensions

The PCECC's procedures and PCEP extensions will be defined in a separate document.

9. Acknowledgments

We would like to thank Robert Tao, Changjing Yan, Tieying Huang for their useful comments and suggestions.

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

10.2. Informative References

- [RFC5441] Vasseur, JP., Zhang, R., Bitar, N., and JL. Le Roux, "A Backward-Recursive PCE-Based Computation (BRPC) Procedure to Compute Shortest Constrained Inter-Domain Traffic Engineering Label Switched Paths", RFC 5441, April 2009.
- [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.
- [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-07 (work in progress), October 2013.
- [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-03 (work in progress), October 2013.
- [I-D.ali-pce-remote-initiated-gmpls-lsp] Ali, Z., Sivabalan, S., Filsfils, C., Varga, R., Lopez, V., and O. Dios, "Path Computation Element

- Communication Protocol (PCEP) Extensions for remote-initiated GMPLS LSP Setup", draft-ali-pce-remote-initiated-gmpls-lsp-01 (work in progress), July 2013.
- [I-D.previdi-isis-segment-routing-extensions] Previdi, S., Filsfils, C., Bashandy, A., Gredler, H., and S. Litkowski, "IS-IS Extensions for Segment Routing", draft-previdi-isis-segment-routing-extensions-03 (work in progress), October 2013.
- [I-D.psenak-ospf-segment-routing-extensions] Psenak, P., Previdi, S., Filsfils, C., Gredler, H., Shakir, R., and W. Henderickx, "OSPF Extensions for Segment Routing", draft-psenak-ospf-segment-routing-extensions-03 (work in progress), October 2013.
- [I-D.sivabalan-pce-segment-routing] Sivabalan, S., Medved, J., Filsfils, C., Crabbe, E., and R. Raszuk, "PCEP Extensions for Segment Routing", draft-sivabalan-pce-segment-routing-02 (work in progress), October 2013.
- [I-D.li-mpls-global-label-usecases] Li, Z., Zhao, Q., and T. Yang, "Usecases of MPLS Global Label", draft-li-mpls-global-label-usecases-00 (work in progress), July 2013.

Authors' Addresses

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

EMail: quintin.zhao@huawei.com

Katherine Zhao
Huawei Technologies
2330 Central Expressway
Santa Clara, CA 95050
USA

EMail: Katherine.zhao@huawei.com

Zhenbin Li
Huawei Technologies
Huawei Bld., No.156 Beiqing Rd.
Beijing 100095
China

EMail: lizhenbin@huawei.com

Zekung Ke
Tencent Holdings Ltd.
Shenzhen
China

EMail: kinghe@tencent.com

