

PCE Working Group
Internet-Draft
Intended status: Standards Track
Expires: January 1, 2015

H. Ananthakrishnan
Juniper Networks
S. Sivabalan
Cisco
C. Barth
R. Torvi
Juniper Networks
I. Minei
E. Crabbe
Google, Inc
June 30, 2014

PCEP Extensions for MPLS-TE LSP Path Protection with stateful PCE
draft-ananthakrishnan-pce-stateful-path-protection-00.txt

Abstract

A stateful Path Computation Element (PCE) is capable of computing as well as controlling via Path Computation Element Protocol (PCEP) Multiprotocol Label Switching Traffic Engineering Label Switched Paths (MPLS LSP). Furthermore, it is also possible for a stateful PCE to create, maintain, and delete LSPs. This document describes PCEP extension to associate two or more LSPs to provide end-to-end path protection.

Status of This Memo

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

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

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

This Internet-Draft will expire on January 1, 2015.

Copyright Notice

Copyright (c) 2014 IETF Trust and the persons identified as the document authors. All rights reserved.

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

Table of Contents

1. Introduction	2
2. Terminology	3
3. PCEP Extensions	4
4. Operation	5
4.1. State Synchronization	5
4.2. Error Handling	5
5. IANA considerations	6
5.1. Association Type	6
5.2. PCEP Errors	6
6. Security Considerations	6
7. References	6
7.1. Normative References	6
7.2. Information References	7

1. Introduction

[RFC5440] describes PCEP for communication between a Path Computation Client (PCC) and a PCE or between one a pair of PCEs. A PCE computes paths for MPLS-TE LSPs based on various constraints and optimization criteria.

Stateful pce [I-D.ietf-pce-stateful-pce] specifies a set of extensions to PCEP to enable stateful control of paths such as MPLS 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. Furthermore, a mechanism to dynamically instantiate LSPs on a PCC based on the requests from a stateful PCE or a controller using stateful PCE is specified in [I-D.ietf-pce-pce-initiated-lsp].

Path protection refers to a paradigm in which the working LSP is protected by one or more protection LSP(s). When the working LSP fails, protection LSP(s) is/are activated. When the working LSPs are

computed and controlled by the PCE, there is benefit in a mode of operation where protection LSPs are as well.

This document specifies a stateful PCEP extension to associate two or more LSPs for the purpose of setting up path protection. The proposed extension covers the following scenarios:

1. A protection LSP is initiated on a PCC by a stateful PCE which retains the control of the LSP. The PCE is responsible for computing the path of the LSP and updating the PCC with the information about the path.
2. A PCC initiates a protection LSP and retains the control of the LSP. The PCC computes the path and updates the PCE with the information about the path as long as it controls the LSP.
3. A PCC initiates a protection LSP and delegates the control of the LSP to a stateful PCE. The PCE may compute the path for the LSP and update the PCC with the information about the path as long as it controls the LSP.

Note that protection LSP can be established (e.g., using RSVP-TE signaling) prior to the failure (in which case the LSP is said to be in standby mode) or post failure of the corresponding working LSP according to the operator choice/policy.

2. Terminology

The following terminologies are used in this document:

AGID: Association Group ID.

ERO: Explicit Route Object.

LSP: Label Switched Path.

PCC: Path Computation Client.

PCE: Path Computation Element

PCEP: Path Computation Element Protocol.

PPAG: Path Protection Association Group.

TLV: Type, Length, and Value.

3. PCEP Extensions

LSPs are not associated by listing the other LSPs with which they interact, but rather by making them belong to an association group referred to as "Path Protection Association Group" (PPAG) in this document. All LSPs join a PPAG individually. PPAG is based on the generic Association object used to associate two or more LSPs specified in [I-D.minei-pce-association-group]. A member of a PPAG can take the role of working or protection LSP. This document defines a new association type called "Path Protection Association Type" of value TBD. A PPAG can have one working LSP and one or more protection LSPs. The source and destination of all LSPs within a PPAG MUST be the same.

The format of the Association object used for PPAG is shown in Figure 1:

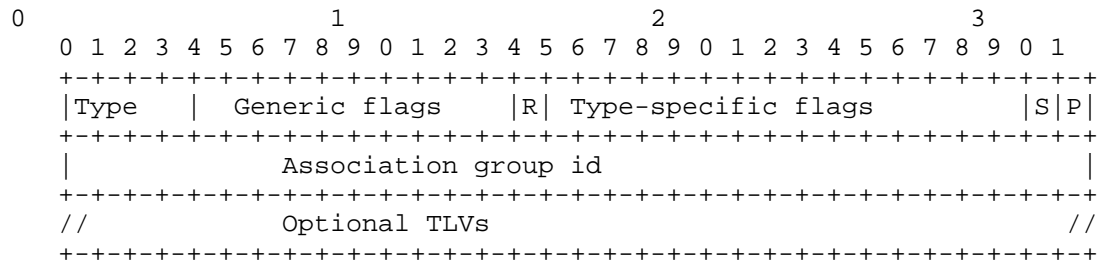


Figure 1: The Association Object format

Type - TBD for the Path Protection Association Type

The description of the flags are as follows:

The 'P' Flag indicates whether the LSP associated with the PPAG is working or protection LSP. If this flag is set, the LSP is protection LSP.

The 'S' Flag if P flag is set, S flag indicates whether the protection LSP associated with the PPAG is in standby mode (e.g., signaled via RSVP-TE prior to failure). The S flag is ignored if P flag is set to 0.

4. Operation

A PCE can create/update working and protection LSPs independently. However, it can add a protection LSP to a PPAG only after adding a working LSP to that group. As specified in [I-D.minei-pce-association-group], Association Group ID (AGID) is allocated by PCC. In order to reserve an AGID, PCE sends an association object with AGID of 0 either in PCInitiate message or PCUpd message for a working LSP, with both the P and S flags set to 0. Upon receiving an association object with AGID of 0, PCC MUST allocate a new AGID and send it the PCE via PCRpt message. Once the PCE receives the AGID, it can either create one or more protection LSP(s) and add it/them to the PPAG or simply add already existing LSP(s) to the PPAG.

A PCE can remove a protection LSP from a PPAG as specified in [I-D.minei-pce-association-group].

A PCC can associate a set of LSPs under its control for path protection purpose. Similarly, the PCC can remove one or more LSPs under its control from the corresponding PPAG. In both cases, the PCC must report the change in association to PCE(s) via PCRpt message.

The forwarding behavior after failure of the protected LSP, in particular how and whether traffic will be load balanced among protection paths will be detailed in a future version of this document.

4.1. State Synchronization

During state synchronization, a PCC MUST report all the existing path protection association groups as well as any path protection flags to PCE(s). Following the state synchronization, the PCE MUST remove all stale path protection associations.

4.2. Error Handling

All LSPs (working or protection) within a PPAG MUST have the same source and destination. If a PCE attempts to add an LSP to a PPAG and the source and/or destination of the LSP is/are different from the LSP(s) in the PPAG, the PCC MUST send PCErr with Error-Type= TBD (Path Protection Association Error) and Error-Value = 1 (End points mismatch).

5. IANA considerations

5.1. Association Type

This document defines a new association type for path protection as follows:

Association Type Value	Association Name	Reference
1	Path Protection Association	This document

5.2. PCEP Errors

This document defines new Error-Type and Error-Value related to path protection association as follows:

Error-Type	Meaning
25	Path Protection Association error: Error-value=1: End-Points mismatch

6. Security Considerations

The same security considerations apply in head end as described in [I-D.ietf-pce-pce-initiated-lsp]

7. References

7.1. Normative References

[I-D.ietf-pce-pce-initiated-lsp]
Crabbe, E., Minei, I., Sivabalan, S., and R. Varga, "PCEP Extensions for PCE-initiated LSP Setup in a Stateful PCE Model", draft-ietf-pce-pce-initiated-lsp-01 (work in progress), June 2014.

[I-D.ietf-pce-stateful-pce]
Crabbe, E., Minei, I., Medved, J., and R. Varga, "PCEP Extensions for Stateful PCE", draft-ietf-pce-stateful-pce-09 (work in progress), June 2014.

- [I-D.minei-pce-association-group]
Minei, I., Crabbe, E., Sivabalan, S., Ananthakrishnan, H., Zhang, X., and Y. Tanaka, "PCEP Extensions for establishing relationships between sets of LSPs", draft-minei-pce-association-group-00 (work in progress), June 2014.
- [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.
- [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.

7.2. Information References

- [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.
- [RFC5394] Bryskin, I., Papadimitriou, D., Berger, L., and J. Ash, "Policy-Enabled Path Computation Framework", RFC 5394, December 2008.
- [RFC5557] Lee, Y., Le Roux, JL., King, D., and E. Oki, "Path Computation Element Communication Protocol (PCEP) Requirements and Protocol Extensions in Support of Global Concurrent Optimization", RFC 5557, July 2009.

Authors' Addresses

Hariharan Ananthakrishnan
Juniper Networks
1194 N Mathilda Ave,
Sunnyvale, CA, 94086
USA

EMail: hanantha@juniper.net

Siva Sivabalan
Cisco
2000 Innovation Drive
Kanata, Ontario K2K 3E8
Canada

EMail: msiva@cisco.com

Colby Barth
Juniper Networks
1194 N Mathilda Ave,
Sunnyvale, CA, 94086
USA

EMail: cbarth@juniper.net

Raveendra Torvi
Juniper Networks
1194 N Mathilda Ave,
Sunnyvale, CA, 94086
USA

EMail: rtorvi@juniper.net

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

EMail: inaminei@google.com

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

EMail: edc@google.com

PCE Working Group
Internet-Draft
Intended status: Informational
Expires: April 13, 2015

D. Dhody
Huawei Technologies
October 10, 2014

Informal Survey into Include Route Object (IRO) Implementations in Path
Computation Element communication Protocol (PCEP)
draft-dhody-pce-iro-survey-00

Abstract

During discussions of a document to provide a standard representation and encoding of Domain-Sequence within the Path Computation Element (PCE) communication Protocol (PCEP) for communications between a Path Computation Client (PCC) and a PCE, or between two PCEs. It was determined that there was a need for clarification with respect to the ordered nature of the Include Route Object (IRO).

Since there was a proposal to have a new IROtype with ordering, as well as handling of Loose bit, it felt necessary to conduct a survey of the existing and planned implementations.

This document summarizes the survey questions and captures the results. Some conclusions are also presented.

This survey was informal and conducted via email. Responses were collected and anonymized by the PCE working group chairs.

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, 2015.

Copyright Notice

Copyright (c) 2014 IETF Trust and the persons identified as the document authors. All rights reserved.

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

Table of Contents

1. Introduction	2
2. Survey Details	3
2.1. Survey Preamble	3
2.2. Survey Questions	3
3. Respondents	5
4. Results	5
5. Conclusions	7
5.1. Proposed Action	7
6. Security Considerations	8
7. IANA Considerations	8
8. Acknowledgments	8
9. References	8
9.1. Normative References	8
9.2. Informative References	8
Appendix A. Contributor Addresses	10

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.

[RFC5440] defines the Include Route Object (IRO) to specify that the computed path must traverse a set of specified network elements. The specification did not mention if IRO is an ordered or un-ordered list of sub-objects. It mentioned that the L bit (loose) has no meaning within an IRO.

[RFC5441] suggested the use of IRO to indicate the sequence of domains to be traversed during inter-domain path computation.

During discussion of [I-D.ietf-pce-pcep-domain-sequence] it was proposed to have a new IRO type with ordered nature, as well as handling of L bit.

In order to discover the current state of affairs amongst implementations a survey of the existing and planned implementations was conducted. This survey was informal and conducted via email. Responses were collected and anonymized by the PCE working group chair.

This document summarizes the survey questions and captures the results. Some conclusions are also presented.

2. Survey Details

2.1. Survey Preamble

The survey was introduced with the following text.

Hi PCE WG.

To address the issues associated with draft-ietf-pce-pcep-domain-sequence and "Include Route Object" in PCEP, Dhruv has proposed to start a small survey. If implementers agree that we need to clarify this, they would be much welcome to answer the attached questions.

Dhruv will process the results, but to improve confidentiality, answers may be sent privately to the chairs.

Thanks,

JP & Julien, on behalf of Dhruv

2.2. Survey Questions

The following survey questions were asked, the survey questionnaire is listed verbatim below.

During discussion of draft-ietf-pce-pcep-domain-sequence-05, it has been noted that RFC 5440 does not define whether the sub-objects in the IRO are ordered or unordered.

We would like to do an informal and *confidential* survey of current implementations, to help clarify this situation.

1. IRO Encoding

- a. Does your implementation construct IRO?
 - b. If your answer to part (a) is Yes, does your implementation construct the IRO as an ordered list always, sometimes or never?
 - c. If your answer to part (b) is Sometimes, what criteria do you use to decide if the IRO is an ordered or unordered list?
 - d. If your answer to part (b) is Always or Sometimes, does your implementation construct the IRO as a sequence of strict hops or as a sequence of loose hops?
2. IRO Decoding
 - a. Does your implementation decode IRO?
 - b. If your answer to part (a) is Yes, does your implementation interpret the decoded IRO as an ordered list always, sometimes or never?
 - c. If your answer to part (b) is Sometimes, what criteria do you use to decide if the IRO is an ordered or unordered list?
 - d. If your answer to part (b) is Always or Sometimes, does your implementation interpret the IRO as a sequence of strict hops or as a sequence of loose hops?
3. Impact
 - a. Will there be an impact to your implementation if RFC 5440 is updated to state that the IRO is an ordered list?
 - b. Will there be an impact to your implementation if RFC 5440 is updated to state that the IRO is an unordered list?
 - c. If RFC 5440 is updated to state that the IRO is an ordered list, will there be an impact to your implementation if RFC 5440 is also updated to allow IRO sub-objects to use the loose bit (L-bit)?
4. Respondents
 - a. Are you a Vendor/Research Lab/Software House/Other (please specify)?

- b. If your answer to part (a) is Vendor, is the implementation for a shipping product, product under development or a prototype?

3. Respondents

Total 9 responses were received from vendors, software houses, and research labs. Vendors made responses for their current shipping products as well as products that they currently have under development.

- o Total Number of Respondents: 9
 - * Vendors: 4
 - + Shipping Product: 1
 - + Product Under Development: 1
 - + Prototype: 1
 - + Unknown: 1
 - * Software House: 1
 - * Research Labs: 2
 - + Operator's Research Facility: 1
 - * Open Source: 1
 - + Shipped Release: 1
 - * Others (or Unknown): 1

4. Results

	Questions	Response
1a	Does your implementation construct IRO?	yes (9)
1b	Does your implementation construct the IRO as an ordered list always, sometimes or never?	always (8), never (1)
1c	What criteria do you use to decide if the IRO is an ordered or unordered list?	none (9)
1d	Does your implementation construct the IRO as a sequence of strict hops or as a sequence of loose hops?	strict (5), loose (2), both (2)

Table 1: IRO Encoding

Regarding IRO encodings, most implementations construct IRO in an ordered fashion and consider it to be an ordered list. More than half of implementation under survey consider the IRO sub-objects as strict hops, others consider loose or support both.

	Questions	Response
2a	Does your implementation decode IRO?	yes (9)
2b	Does your implementation interpret the decoded IRO as an ordered list always, sometimes or never?	always (7), sometimes (1), never (1)
2c	What criteria do you use to decide if the IRO is an ordered or unordered list?	none (9)
2d	Does your implementation interpret the IRO as a sequence of strict hops or as a sequence of loose hops?	strict (5), loose (2), both (2)

Table 2: IRO Decoding

Regarding IRO decoding, most implementations interpret IRO as an ordered list. More than half of implementation under survey consider the IRO sub-objects as strict hops, others consider loose or support both.

	Questions	Response
3a	Will there be an impact to your implementation if [RFC5440] is updated to state that the IRO is an ordered list?	none (9)
3b	Will there be an impact to your implementation if [RFC5440] is updated to state that the IRO is an unordered list?	yes (5), no (4)
3c	will there be an impact to your implementation if [RFC5440] is also updated to allow IRO sub-objects to use the loose bit (L-bit)?	none (5), yes(1), yes-but-small (3)

Table 3: Impact

It is interesting to note that most implementation that responded to the survey finds that there is no impact to their existing or under-development implementation if [RFC5440] is updated to state that the IRO as an ordered list. Further most implementations find that support for loose bit (L-bit) for IRO has minimal or no impact on their implementation.

5. Conclusions

The results shown in this survey seems to suggest that most implementations would be fine with updating [RFC5440] to specify IRO as an ordered list with no impact on the shipping or under-development products. It is also the conclusion of this survey to suggest that it would be helpful to update [RFC5440] to enable support for loose bit (L-bit) such that both strict and loose hops could be supported in the IRO.

5.1. Proposed Action

The proposed action is as follows:

- o Update [RFC5440] to specify IRO as an ordered list.
- o Update [RFC5440] to specify support for loose bit (L-bit) for IRO.
- o Remove the new IRO option from draft-ietf-pce-pcep-domain-sequence-05.

An update to draft-ietf-pce-pcep-domain-sequence-05 is one possible way to handle all of the above proposed action points.

6. Security Considerations

This survey defines no protocols or procedures and so includes no security-related protocol changes. Clarification in the supported IRO ordering will not have any negative security impact. The survey responses in this document were collected by email and that email was not authenticated, although responses were sent to the respondents that might have triggered alarms if the responses were spoofed. Spoofed or malicious responses could represent an attack on the IETF process and so this survey should be treated with some caution where there is reason to suspect such an attack. Further, this survey was compiled and anonymized by the working group chairs.

7. IANA Considerations

This informational document makes no requests to IANA for action.

8. Acknowledgments

A special thanks to author of [I-D.farrel-ccamp-ero-survey], this document borrow some of the structure and text from it.

9. References

9.1. Normative References

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

9.2. 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.

[I-D.ietf-pce-pcep-domain-sequence]
Dhody, D., Palle, U., and R. Casellas, "Standard Representation Of Domain-Sequence", draft-ietf-pce-pcep-domain-sequence-05 (work in progress), July 2014.

[I-D.farrel-ccamp-ero-survey]

Farrel, A., "Informal Survey into Explicit Route Object
Implementations in Generalized Multiprotocol Labels
Switching Signaling Implementations", draft-farrel-ccamp-
ero-survey-00 (work in progress), May 2006.

Appendix A. Contributor Addresses

Julien Meuric
Orange

EMail: julien.meuric@orange.com

Jean Philippe Vasseur
Cisco Systems, Inc.

EMail: jpv@cisco.com

Jonathan Hardwick
Metaswitch
100 Church Street
Enfield EN2 6BQ
UK

EMail: jonathan.hardwick@metaswitch.com

Author's Address

Dhruv Dhody
Huawei Technologies
Leela Palace
Bangalore, Karnataka 560008
INDIA

EMail: dhruv.ietf@gmail.com

PCE Working Group
Internet-Draft
Intended status: Informational
Expires: January 5, 2015

D. Dhody
X. Zhang
Huawei Technologies
July 4, 2014

Stateful Path Computation Element (PCE) Inter-domain Considerations
draft-dhody-pce-stateful-pce-interdomain-00

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 path calculations for its associated Path Computation Clients (PCCs). Furthermore, PCEs are used to compute shortest constrained Traffic Engineering Label Switched Paths (TE LSPs) in Multiprotocol Label Switching (MPLS) and Generalized MPLS (GMPLS) networks across multiple domains.

This document describes general considerations for the deployment of stateful PCE(s) in inter-domain scenarios including inter-area and inter-AS.

Status of This Memo

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

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

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

This Internet-Draft will expire on January 5, 2015.

Copyright Notice

Copyright (c) 2014 IETF Trust and the persons identified as the document authors. All rights reserved.

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

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

Table of Contents

1. Introduction	2
2. Overview	3
2.1. LSP State Synchronization	4
3. Stateful PCE Deployments	4
3.1. Single Stateful PCE, Multiple Domains	4
3.2. Multiple Stateful PCE, Multiple Domains	5
3.2.1. Per Domain Path Computation	6
3.2.2. Backward-Recursive PCE-based Computation	7
3.2.3. Hierarchical PCE	7
4. Other Considerations	8
4.1. Delegation	8
5. Security Considerations	9
6. Manageability Considerations	9
6.1. Control of Function and Policy	9
6.2. Information and Data Models	9
6.3. Liveness Detection and Monitoring	9
6.4. Verify Correct Operations	9
6.5. Requirements On Other Protocols	9
6.6. Impact On Network Operations	10
7. IANA Considerations	10
8. Acknowledgments	10
9. References	10
9.1. Normative References	10
9.2. Informative References	10
Appendix A. Contributor Addresses	12

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-app] describes general considerations for a stateful PCE deployment and examines its applicability and benefits, as well as its challenges and limitations through a number of use cases. [I-D.ietf-pce-stateful-pce] describes a set of extensions to PCEP to provide stateful control. A stateful PCE has access to not only the information carried by the network's Interior

Gateway Protocol (IGP), but also the set of active paths and their reserved resources for its computations. The additional state allows the PCE to compute constrained paths while considering individual LSPs and their interactions.

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

This document presents general considerations for stateful PCE(s) deployment in multi-domain scenarios.

2. Overview

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. 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. [I-D.ietf-pce-stateful-pce] applies equally to MPLS-TE and GMPLS LSPs and distinguishes between an active and a passive stateful PCE. A passive stateful PCE uses LSP state information to optimize path computations but does not actively update LSP state. In contrast, an active stateful PCE may issue recommendations to the network. For example, an active stateful PCE may update LSP parameters for those LSPs that have been delegated, by its PCCs, the control over to the PCE.

The capability to compute the routes of end-to-end inter-domain MPLS-TE LSPs is expressed as requirements in [RFC4105] and [RFC4216] and may be realized by PCE(s). PCEs may use one of the following mechanisms to compute end-to-end paths:

- o a per-domain path computation technique [RFC5152];
- o a Backward-Recursive PCE-based Computation (BRPC) mechanism [RFC5441];
- o a Hierarchical PCE mechanism [RFC6805];

This document examines the stateful PCE inter-domain considerations for all of these mechanisms.

2.1. LSP State Synchronization

The population of the LSP-DB using information received from PCCs (ingress LSR) is supported by the stateful PCE extensions defined in [I-D.ietf-pce-stateful-pce] , i.e., via LSP state report messages.

The inter-domain LSP state is synchronised to the ingress-PCE from the ingress LSR (PCC), but this PCC cannot synchronise to other PCEs (in transit or egress domains), thus other mechanism must be investigated for this purpose.

3. Stateful PCE Deployments

There are multiple models to perform PCE-based inter-domain path computation:

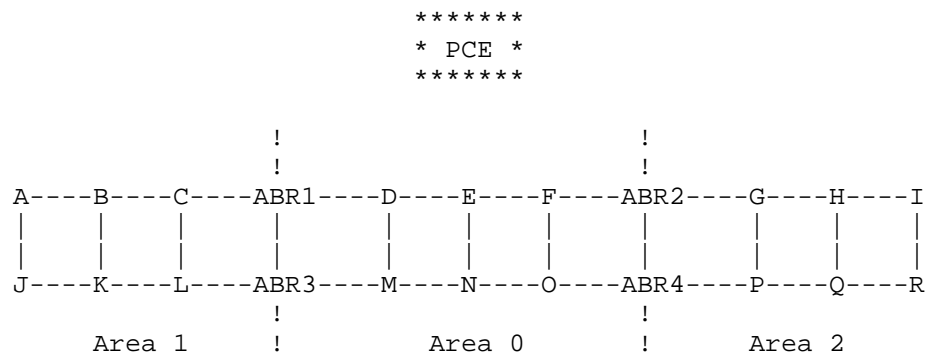
- o A single PCE;
- o Multiple PCE;
 - * without inter-PCE communication;
 - * with inter-PCE communication;

This section describe stateful PCE considerations for each of these deployment models.

3.1. Single Stateful PCE, Multiple Domains

In this model, inter-domain path computation is performed by a single stateful PCE that has topology visibility into all domains. The inter-domain LSP state is synchronised to this PCE from the ingress LSR (PCC) itself. This PCC may also choose to delegate control over this LSP to the PCE. Thus this model is similar to a single domain in all aspects.

Following figure show an example of inter-area case comprising of Area 0,1 and 2. A single stateful PCE is deployed for all areas.



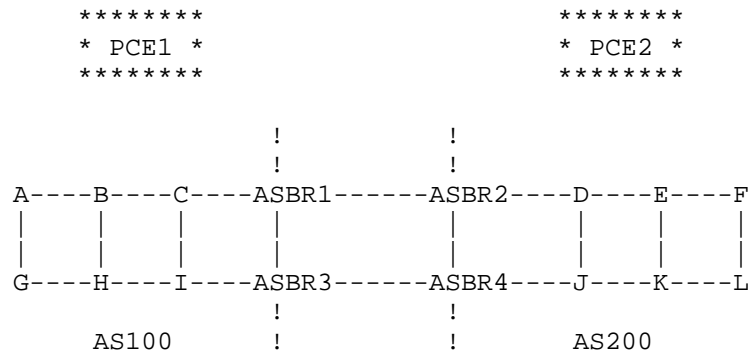
In this model PCE has visibility into the topology of all domains as well as the state of all active LSPs including inter-domain LSPs. This model is thus well suited to take advantage of all stateful PCE capabilities.

It should be noted that a single PCE may not be possible because of administrative and confidentiality concerns.

3.2. Multiple Stateful PCE, Multiple Domains

In this model, there is at least one PCE per domain, and each PCE has topology visibility restricted to its own domain. The inter-domain LSP state is synchronised to the ingress-PCE from the ingress LSR (PCC), but this PCC cannot synchronise to other PCEs (in transit or egress domains). This PCC may also choose to delegate control over this LSP to the Ingress-PCE, which may issue inter-domain path computation or re-optimization request to other PCEs. An inter-domain LSP that originates in the domain, is synchronised to the PCE in that domain. But a mechanism is needed to synchronize state of inter-domain LSP that do not originate in the domain. In other words, inter-domain LSP state should also be synchronised to transit and egress PCEs.

Following figure show an example of inter-AS case comprising of AS 100 and AS 200. A stateful PCE is deployed per AS.



In order to conceal the information, a PCE may use path-key based confidentiality mechanisms as per [RFC5520].

This section further describes considerations with respect to each of the inter-domain path computation techniques.

3.2.1. Per Domain Path Computation

The per domain path computation technique [RFC5152] is based on Multiple PCE Path Computation without Inter-PCE Communication Model as described in [RFC4655]. It 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.

The ingress LSR would synchronise the the state to the ingress PCE, further the entry boundary nodes should synchronize the state of inter-domain LSP to transit and egress PCEs. Note that the BN on the path of an LSP can probably see the path (through the Record Route object in RSVP-TE signaling [RFC3209]) and knows the bandwidth reserved for the LSP. Thus each entry BN along the path could be made responsible to synchronise the LSP state to the transit/egress PCE(s).

Since the stateful PCE(s) do not communicate during this inter-domain path computation technique and each entry BN would perform path computation via Path Computation Request (PCReq) and Reply (PCRep) messages, a passive stateful PCE is well suited for this case.

In case of delegation to the ingress PCE (active stateful PCE), it would be capable of loose path computation only and make updates to the ingress LSR with this limited visibility. The entry BN would perform path computation via Path Computation Request and Reply

messages (and thus rely on the passive stateful mode). Thus the inter-domain LSP is delegated only to the ingress PCE.

3.2.2. Backward-Recursive PCE-based Computation

The BRPC [RFC5441] technique is based on Multiple PCE Path Computation with Inter-PCE Communication Model as described in [RFC4655]. It 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 may be known before the path computation, but it can also be used when the domain path is unknown and determined during path computation.

As described in Section 3.2.1, the entry boundary nodes may synchronize the state of inter-domain LSPs to transit and egress PCEs. An alternative approach may be for each PCE to synchronise the state along the path across domains, i.e., each PCE would report the state to the next PCE(s) in the adjacent domain along the domain sequence of the inter-domain path. A mechanism similar to LSP-DB backup [I-D.palle-pce-stateful-pce-lspdb-sync] may be utilized for this purpose.

Some path segment in the end to end path may also be hidden via path-key as per [RFC5520] during state synchronization.

In case of passive path computation request to the ingress PCE from the ingress LSR the BRPC path computation procedure is applied to compute end-to-end path by using PCReq and PCRep messages among stateful PCE(s) in passive mode.

In case of delegation to the ingress PCE (active stateful PCE), the ingress PCE may trigger the end-to-end path computation via the same BRPC procedure using the path computation request and reply messages among stateful PCE(s) in passive mode. For re-optimization or update the ingress PCE still rely on the same BRPC procedure triggered by the ingress PCE. Ultimately the inter-domain LSP is delegated to the ingress PCE and only the ingress PCE can issue updates to the inter-domain LSP. It may trigger E2E path re-optimization with help of transit/egress PCE using the BRPC procedure.

3.2.3. 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.

As described in Section 3.2.1 and Section 3.2.2, the entry boundary nodes may synchronize the state of inter-domain LSP to transit and egress child PCEs. If the parent PCE provides the sequence of domains and BRPC procedure is used to get the E2E path, each PCE may be responsible to synchronise the state along the path across domains similar to Section 3.2.2. An alternative approach may be for ingress PCE to synchronise LSP state with the Parent PCE and it may further synchronise the state to the child PCE(s) along the path across domains, i.e. parent PCE would report the state to the child PCE(s) along the domain sequence.

Some path segment in the end to end path may also be hidden via path-key as per [RFC5520] during state synchronization.

In case of passive path computation request to the ingress PCE from the ingress LSR, the H-PCE path computation procedure is applied to compute sequence of domains or end-to-end path by using PCReq and PCRep messages among stateful PCE(s) in passive mode.

In case of delegation to the ingress PCE (active stateful PCE), the ingress PCE may trigger the H-PCE path computation via the same procedure using the PCReq and PCRep messages among stateful PCE(s) in passive mode. For re-optimization or update the ingress PCE still rely on the same H-PCE procedure triggered by the ingress PCE. Ultimately the inter-domain LSP is delegated to the ingress PCE and only the ingress PCE can issue updates to the inter-domain LSP. It may trigger E2E path re-optimization with help of parent and child PCEs using the H-PCE procedure.

4. Other Considerations

4.1. Delegation

As noted in this document, the inter-domain LSP is delegated to the ingress PCE and only the ingress PCE can issue updates to the inter-domain LSP. The ingress PCE is responsible to trigger E2E path re-optimization.

Thus the ingress PCE can recommend updation for all aspects of the inter-domain LSP including the segment of path in another domain (which it may have computed with the help of other cooperating PCEs). These interaction between PCEs for the inter-domain path computation are done using PCReq/PCRep messages (i.e., in a passive mode).

The transit/egress PCE cannot update any attribute of the inter-domain LSP on its own as it may not have any interaction with the ingress LSR. A mechanism may be developed for transit/egress PCE to inform the ingress PCE to trigger E2E re-optimization and choose to

update the inter-domain LSP based on the result. Also the ingress PCE may use combination of local information and events along with some external mechanism (management / monitoring interface) to trigger E2E path re-optimization.

Though Ingress PCE can recommend update for path segments in other domains, the entry boundary node of that domain can apply policy control during signalling as explained in [RFC4105] and [RFC4216].

5. Security Considerations

The security considerations are as per [RFC5440] and [I-D.ietf-pce-stateful-pce]. Any multi-domain operation necessarily involves the exchange of information across domain boundaries. This may represent a significant security and confidentiality risk especially when the domains are controlled by different commercial entities. PCEP allows individual PCEs to maintain confidentiality of their domain path information by using path-keys [RFC5520].

6. Manageability Considerations

6.1. Control of Function and Policy

Mechanisms defined in this document do not imply any new control of function and policy requirements.

6.2. Information and Data Models

[I-D.ietf-pce-pcep-mib] describes the PCEP MIB, there are no new MIB Objects for this document.

6.3. Liveness Detection and Monitoring

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

6.4. Verify Correct Operations

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

6.5. Requirements On Other Protocols

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

6.6. Impact On Network Operations

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

7. IANA Considerations

This is an informational document and has no IANA considerations.

8. Acknowledgments

TBD.

9. References

9.1. Normative References

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

[I-D.ietf-pce-stateful-pce]
Crabbe, E., Minei, I., Medved, J., and R. Varga, "PCEP Extensions for Stateful PCE", draft-ietf-pce-stateful-pce-09 (work in progress), June 2014.

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.

[RFC4105] Le Roux, J., Vasseur, J., and J. Boyle, "Requirements for Inter-Area MPLS Traffic Engineering", RFC 4105, June 2005.

[RFC4216] Zhang, R. and J. Vasseur, "MPLS Inter-Autonomous System (AS) Traffic Engineering (TE) Requirements", RFC 4216, November 2005.

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

[RFC5152] Vasseur, JP., Ayyangar, A., and R. Zhang, "A Per-Domain Path Computation Method for Establishing Inter-Domain Traffic Engineering (TE) Label Switched Paths (LSPs)", RFC 5152, February 2008.

- [RFC5441] Vasseur, JP., Zhang, R., Bitar, N., and JL. Le Roux, "A Backward-Recursive PCE-Based Computation (BRPC) Procedure to Compute Shortest Constrained Inter-Domain Traffic Engineering Label Switched Paths", RFC 5441, April 2009.
- [RFC5520] Bradford, R., Vasseur, JP., and A. Farrel, "Preserving Topology Confidentiality in Inter-Domain Path Computation Using a Path-Key-Based Mechanism", RFC 5520, April 2009.
- [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.
- [I-D.ietf-pce-stateful-pce-app]
Zhang, X. and I. Minei, "Applicability of a Stateful Path Computation Element (PCE)", draft-ietf-pce-stateful-pce-app-02 (work in progress), June 2014.
- [I-D.ietf-pce-pcep-mib]
Koushik, K., Emile, S., Zhao, Q., King, D., and J. Hardwick, "Path Computation Element Protocol (PCEP) Management Information Base", draft-ietf-pce-pcep-mib-08 (work in progress), April 2014.
- [I-D.palle-pce-stateful-pce-lspdb-sync]
Palle, U., Dhody, D., and X. Zhang, "LSP-DB Synchronization between Stateful PCEs", draft-palle-pce-stateful-pce-lspdb-sync-02 (work in progress), January 2014.

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

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

EMail: zhang.xian@huawei.com

Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: January 1, 2015

J. Dong
M. Chen
Huawei Technologies
June 30, 2014

BGP Extensions for Path Computation Element (PCE) Discovery
draft-dong-pce-discovery-proto-bgp-00

Abstract

In network scenarios where Path Computation Element (PCE) is used for centralized path computation, it is desirable for Path Computation Clients (PCCs) to automatically discover the set of PCEs. As BGP has been extended for north-bound distribution of routing and LSP path information to PCE, the PCEs may not participate in Interior Gateway Protocol (IGP) for collecting the routing information, thus the IGP based PCE discovery cannot be used directly in these scenarios. This document specifies the BGP extensions for PCE discovery.

Requirements Language

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

Status of This Memo

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

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

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

This Internet-Draft will expire on January 1, 2015.

Copyright Notice

Copyright (c) 2014 IETF Trust and the persons identified as the document authors. All rights reserved.

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

Table of Contents

1. Introduction	2
2. Carrying PCE Discovery Information in BGP	3
2.1. PCE Address Information	3
2.2. PCE Discovery Attribute	4
3. Operational Considerations	5
4. IANA Considerations	5
5. Security Considerations	5
6. Acknowledgements	5
7. References	5
7.1. Normative References	5
7.2. Informative References	6
Authors' Addresses	6

1. Introduction

In network scenarios where Path Computation Element (PCE) is used for centralized path computation, it is desirable for Path Computation Clients (PCCs) to automatically discover the set of PCEs. As BGP will be used for north-bound distribution of routing and Label Switched Path (LSP) information to PCE[I-D.ietf-idr-ls-distribution] [I-D.ietf-idr-te-lsp-distribution] [I-D.ietf-idr-te-pm-bgp], the PCEs may not participate in Interior Gateway Protocol (IGP) for collecting the routing information, thus the IGP based PCE discovery mechanisms defined in [RFC5088] [RFC5089] cannot be used directly.

This document proposes to extend BGP for PCE discovery in such scenarios. While in each IGP domain, the IGP based PCE discovery mechanism may be used in conjunction with the BGP based PCE discovery. Thus the BGP based PCE discovery is complementary to the existing IGP based mechanisms.

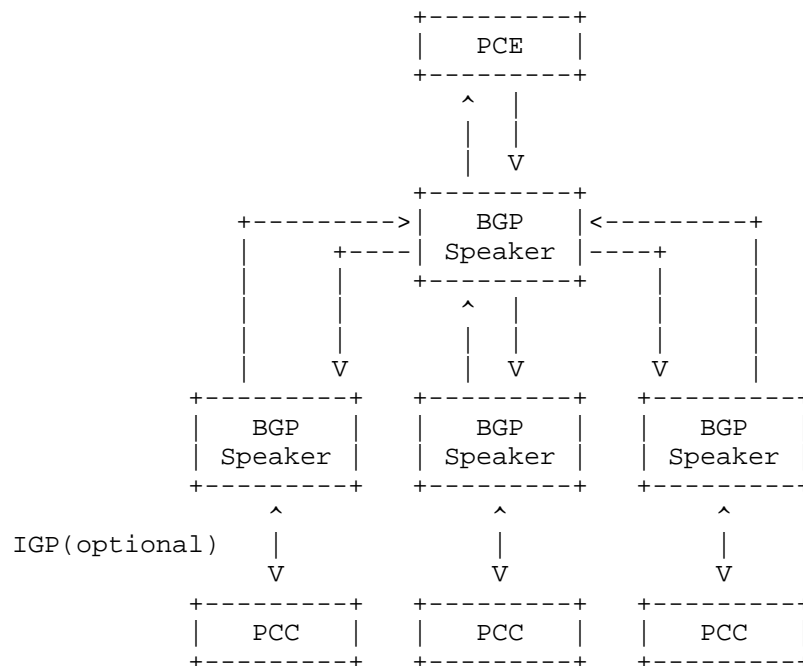


Figure 1. BGP for routing collection and PCE discovery

As shown in the network architecture in Figure 1, BGP is used for both routing information distribution and PCE information discovery. The routing information is distributed from the network elements up to PCE, while the PCE discovery information is advertised from PCE down to PCCs. IGP based PCE discovery mechanism may be used for the distribution of PCE discovery information in each IGP domain.

2. Carrying PCE Discovery Information in BGP

2.1. PCE Address Information

The PCE discovery information is advertised in BGP UPDATE messages using the MP_REACH_NLRI and MP_UNREACH_NLRI attributes [RFC4760]. A new NLRI called PCE_ADDR NLRI is defined for carrying the PCE address information which can be used to reach the PCE. The AFI/SAFI value for the PCE_ADDR NLRI is TBD. In order for two BGP speakers to exchange PCE_ADDR NLRI, they MUST use BGP Capabilities Advertisement [RFC4760] to ensure that both are capable of properly processing such NLRI. This is done by using Capability Code 1 (which indicates Multiprotocol Extensions capabilities), with the AFI/SAFI pair for the PCE_ADDR NLRI.

The format of PCE_ADDR NLRI is shown as below:

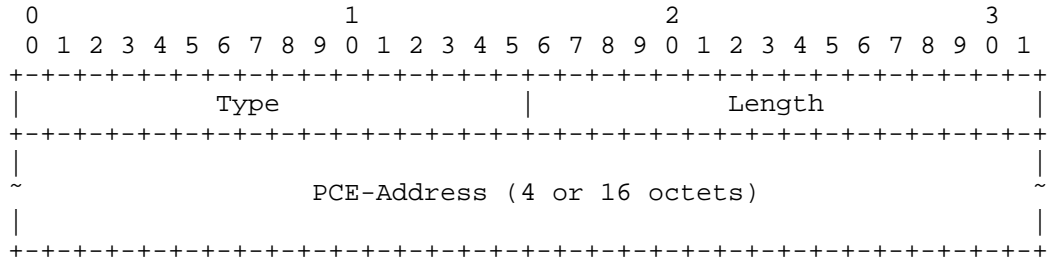


Figure 2. PCE_ADDR NLRI

For PCEs identified by IPv4 address, the Type field SHOULD be set to 1, and the Length field SHOULD be set to 4.

For PCEs identified by IPv6 address, the Type field SHOULD be set to 2, and the Length field SHOULD be set to 16.

2.2. PCE Discovery Attribute

The detailed PCE discovery information is carried in a new optional non-transitive BGP attribute called PCE_DISC Attribute, which consists of a series of PCE Discovery TLVs for specific PCE information. The PCE_DISC attribute SHOULD only be used with PCE_ADDR NLRI.

The format of the PCE Discovery TLV is shown as below:

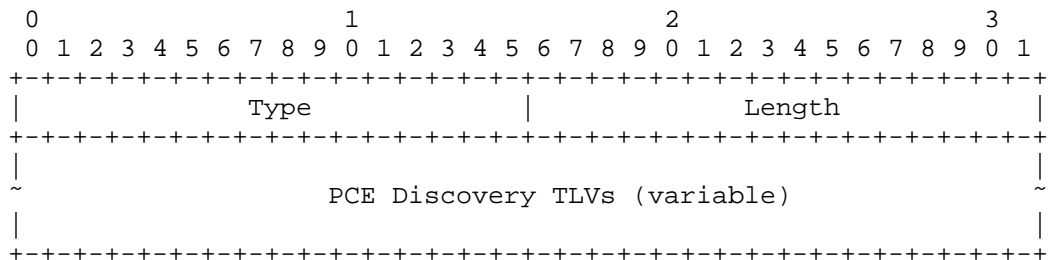


Figure 3. PCE Discovery TLVs

The Type code and format of the PCE Discovery TLVs are consistent with the IGP PCED Sub-TLVs defined in [RFC5088] [RFC5089]. Type 1 is reserved, which is used in IGP based PCE discovery mechanisms to carry PCE Address .

TLV-Type	Length	Name
2	3	PATH-SCOPE TLV
3	variable	PCE-DOMAIN TLV
4	variable	NEIG-PCE-DOMAIN TLV
5	variable	PCE-CAP-FLAGS TLV

The PATH-SCOPE TLV MUST always be carried in the PCE_DISC Attribute. Other TLVs are optional and may facilitate the PCE selection.

More PCE Discovery TLVs may be defined in future.

3. Operational Considerations

Existing BGP operational procedures apply to the advertisement of PCE discovery information. Such information is treated as pure application level data which has no immediate impact on forwarding states.

PCE discovery information is considered relatively stable and does not change frequently, thus this information will not bring significant impact on the amount of BGP updates in the network.

4. IANA Considerations

IANA needs to assign new AFI and SAFI codes for PCE_ADDR NLRI from "Address Family Numbers" and "Subsequent Address Family Identifiers" registry.

IANA needs to assign a new type code for "PCE_DISC" attribute from "BGP Path Attributes" registry.

5. Security Considerations

Procedures and protocol extensions defined in this document do not affect the BGP security model. See [RFC6952] for details.

6. Acknowledgements

The authors would like to thank Zhenbin Li for the discussion and comments.

7. References

7.1. Normative References

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

- [RFC4760] Bates, T., Chandra, R., Katz, D., and Y. Rekhter, "Multiprotocol Extensions for BGP-4", RFC 4760, January 2007.
- [RFC5088] Le Roux, JL., Vasseur, JP., Ikejiri, Y., and R. Zhang, "OSPF Protocol Extensions for Path Computation Element (PCE) Discovery", RFC 5088, January 2008.
- [RFC5089] Le Roux, JL., Vasseur, JP., Ikejiri, Y., and R. Zhang, "IS-IS Protocol Extensions for Path Computation Element (PCE) Discovery", RFC 5089, January 2008.
- [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.

7.2. Informative References

- [I-D.ietf-idr-ls-distribution]
Gredler, H., Medved, J., Previdi, S., Farrel, A., and S. Ray, "North-Bound Distribution of Link-State and TE Information using BGP", draft-ietf-idr-ls-distribution-05 (work in progress), May 2014.
- [I-D.ietf-idr-te-lsp-distribution]
Dong, J., Chen, M., Gredler, H., and S. Previdi, "Distribution of MPLS Traffic Engineering (TE) LSP State using BGP", draft-ietf-idr-te-lsp-distribution-00 (work in progress), January 2014.
- [I-D.ietf-idr-te-pm-bgp]
Wu, Q., Danhua, W., Previdi, S., Gredler, H., and S. Ray, "BGP attribute for North-Bound Distribution of Traffic Engineering (TE) performance Metrics", draft-ietf-idr-te-pm-bgp-00 (work in progress), January 2014.

Authors' Addresses

Jie Dong
Huawei Technologies
Huawei Campus, No. 156 Beiqing Rd.
Beijing 100095
China

Email: jie.dong@huawei.com

Mach(Guoyi) Chen
Huawei Technologies
Huawei Campus, No. 156 Beiqing Rd.
Beijing 100095
China

Email: mach.chen@huawei.com

PCE Working Group
Internet-Draft
Intended status: Standards Track
Expires: April 8, 2018

E. Crabbe
Individual Contributor
I. Minei
Google, Inc.
S. Sivabalan
Cisco Systems, Inc.
R. Varga
Pantheon Technologies SRO
October 5, 2017

PCEP Extensions for PCE-initiated LSP Setup in a Stateful PCE Model
draft-ietf-pce-pce-initiated-lsp-11

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 for stateful PCE provide active 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 8, 2018.

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
2. Terminology	3
3. Architectural Overview	4
3.1. Motivation	4
3.2. Operation Overview	5
4. Support of PCE-initiated LSPs	6
4.1. STATEFUL-PCE-CAPABILITY TLV	6
5. PCE-initiated LSP Instantiation and Deletion	7
5.1. The LSP Initiate Request	7
5.2. The R flag in the SRP Object	8
5.3. LSP Instantiation	9
5.3.1. The Create Flag	11
5.3.2. The SPEAKER-ENTITY-ID TLV	11
5.4. LSP Deletion	12
6. LSP Delegation and Cleanup	12
7. LSP State Synchronization	13
8. Implementation Status	14
9. IANA Considerations	14
9.1. PCEP Messages	14
9.2. LSP Object	15
9.3. SRP object	15
9.4. STATEFUL-PCE-CAPABILITY TLV	15
9.5. PCEP-Error Object	15
10. Security Considerations	16
10.1. Malicious PCE	16
10.2. Malicious PCC	17
11. Acknowledgements	17
12. References	17
12.1. Normative References	17

12.2. Informative References	18
Authors' Addresses	18

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 PCE and PCE, enabling computation of Multiprotocol Label Switching (MPLS) for Traffic Engineering Label Switched Path (TE LSP) characteristics.

[RFC8231] 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

- o mechanisms to effect LSP state synchronization between PCCs and PCEs
- o delegation of control of LSPs to PCEs
- o PCE control of timing and sequence of path computations within and across PCEP sessions

It focuses on a model where LSPs are configured on the PCC and 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 [RFC8051]: Stateful PCE, Delegation.

This document uses the following terms defined in [RFC8231]: Redelegation Timeout Interval, 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 Form (RBNF) encoding as specified in [RFC5511].

3. Architectural Overview

3.1. Motivation

[RFC8231] provides active control over LSPs that are locally configured on the PCC. This model relies on the Label Edge Router (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 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 a software-defined 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 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 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

This document defines the new I flag in the STATEFUL-PCE-CAPABILITY TLV to indicate that the sender supports PCE-initiated LSPs (see details in Section 4.1). A PCC or PCE sets this flag in the Open message during the PCEP Initialization Phase to indicate that it supports the procedures of this document.

This document defines a new PCEP message, the LSP Initiate Request (PCInitiate) message, which a PCE can send to a PCC to request the initiation or deletion of an LSP. The decision when to instantiate or delete a PCE-initiated LSP is out of the scope of this document.

The PCE sends a PCInitiate message to the PCC to request the initiation of an LSP. The PCC creates the LSP using the attributes communicated by the PCE and local values for any unspecified parameters. The PCC generates an LSP State Report (PCRpt) for the LSP, carrying a newly assigned PLSP-ID for the LSP and delegating the LSP to the PCE via the Delegate flag in the LSP object.

The PCE can update the attributes of the LSP by sending subsequent PCUpd messages. Subsequent LSP State Report (PCRpt) and LSP Update Request (PCUpd) messages that the PCC and PCE, respectively, send for the LSP will carry the PCC-assigned PLSP-ID, which uniquely identifies the LSP. See details in Section 5.3.

The PCE sends a PCInitiate message to the PCC to request the deletion of an LSP. To indicate a delete operation, this document defines the new R flag in the SRP object in the PCInitiate message, as described in Section 5.2. As a result of the deletion request, the PCC removes the LSP and sends a PCRpt for the removed state. See details in Section 5.4.

Figure 1 illustrates these message exchanges.

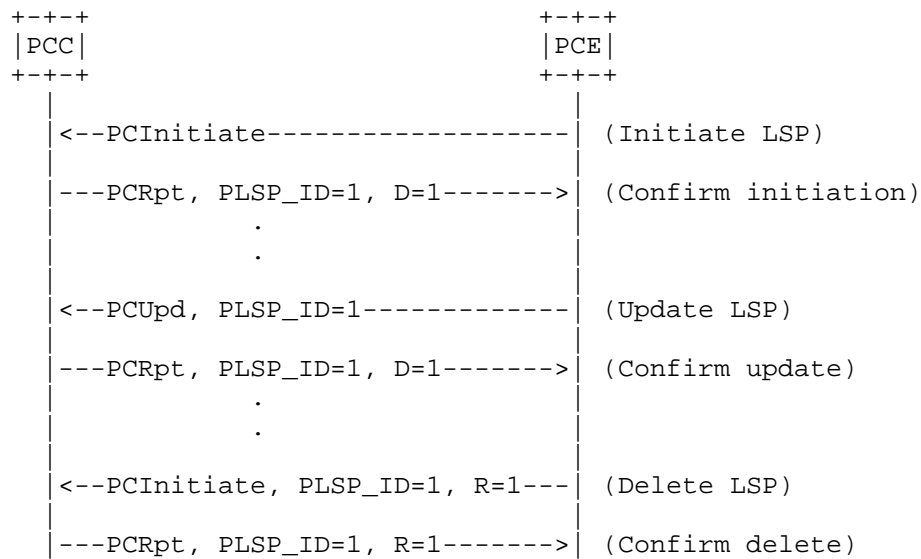


Figure 1: PCE-Initiated LSP lifecycle

4. Support of PCE-initiated LSPs

A PCEP speaker indicates its ability to support PCE-initiated LSPs during the PCEP Initialization phase, as follows. When the PCEP session is created, it sends an Open message with an OPEN object that contains the STATEFUL-PCE-CAPABILITY TLV, defined in [RFC8231]. A new flag, the I (LSP-INSTANTIATION-CAPABILITY) flag, is introduced to this TLV to indicate support for instantiation of PCE-initiated LSPs. A PCE can initiate LSPs only for PCCs that advertised this capability. 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 defined in [RFC8231] and included here for easy reference with the addition of the new I flag.

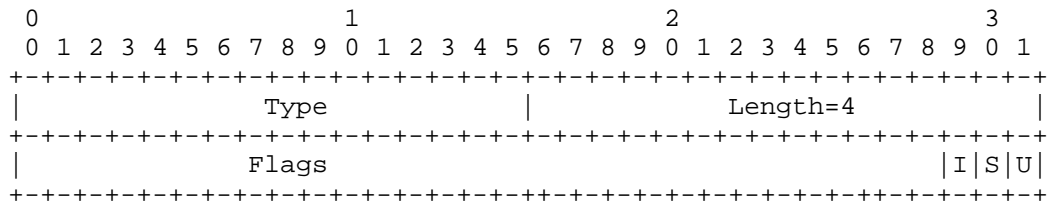


Figure 2: STATEFUL-PCE-CAPABILITY TLV format

A new flag is defined to indicate the sender's support for LSP instantiation by a 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 supports instantiating LSPs. The LSP-INSTANTIATION-CAPABILITY flag must be set by both PCC and PCE in order to enable PCE-initiated LSP instantiation.

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 Request

An LSP Initiate Request (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 12. The PCInitiate message MUST include the SRP and the LSP objects, and MAY contain other objects, as discussed later in this section.

The format of a PCInitiate message is as follows:

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

Where:

<Common Header> is defined in [RFC5440]

```
<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 LSP object is defined in [RFC8231]. The END-POINTS and ERO objects are defined in [RFC5440].

The SRP object is defined in [RFC8231]. The SRP Object contains an SRP-ID-number which is unique within a PCEP session. The PCE increments the last-used SRP-ID-number before it sends each PCInitiate message. The PCC MUST echo the value of the SRP-ID-number in PCErr and PCRpt messages that it sends as a result of the PCInitiate to allow the PCE to correlate them with the corresponding PCInitiate message.

5.2. The R flag in the SRP Object

The format of the SRP object is defined in [RFC8231] and included here for easy reference with the addition of the new R flag.

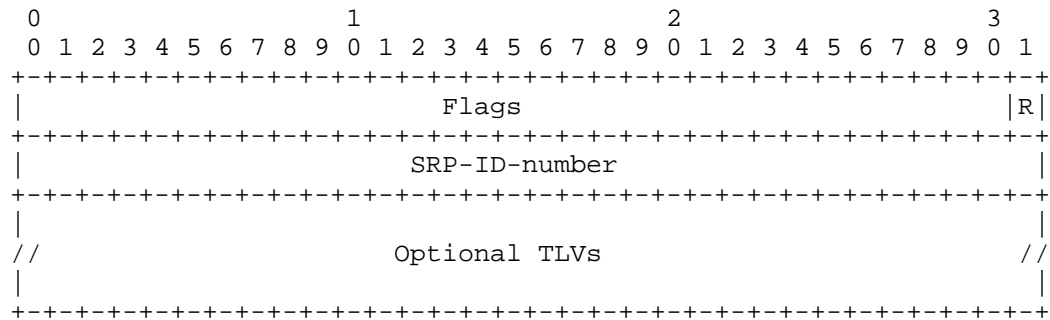


Figure 3: The SRP Object format

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

R (LSP-REMOVE - 1 bit): If set to 0, it indicates a request to create an LSP. If set to 1, it indicates a request to remove an LSP.

5.3. LSP Instantiation

The LSP is instantiated by sending a PCInitiate message. The LSP is set up using RSVP-TE. Extensions for other setup methods are outside the scope of this draft.

The PCInitiate message, when used to instantiate an LSP, MUST contain an LSP object with the reserved PLSP-ID 0. The LSP Object MUST include the SYMBOLIC-PATH-NAME TLV, which is used to correlate between the PCC-assigned PLSP-ID and the LSP.

The PCInitiate message, when used to instantiate an LSP, MUST contain an Explicit Route Object (ERO) for the LSP.

For an instantiation request of an RSVP-signaled LSP, the destination address may be needed. The PCC MAY determine it from a provided object (e.g., ERO) or a local decision. Alternatively, the END-POINTS object MAY be included to explicitly convey the destination addresses to be used in the RSVP-TE signaling. The source address MUST either be specified or left for the PCC to choose by setting it to "0.0.0.0" (if the destination is an IPv4 address) or "::" (if the destination is an IPv6 address).

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 Create flag (Section 5.3.1) set in the LSP object.

If the PCC receives a PCInitiate message with a non-zero PLSP-ID and the R flag in the SRP object set to zero, then it MUST send a PCErr message with Error-type=19 (Invalid Operation) and Error-value=8 (Non-zero PLSP-ID in the PCInitiate message).

If the PCC receives a PCInitiate message without an ERO and the R flag in the SRP object set to zero, then it MUST send a PCErr message with Error-type=6 (Mandatory Object missing) and Error-value=9 (ERO Object missing).

If the PCC receives a PCInitiate message without a SYMBOLIC-PATH-NAME TLV, then it MUST send a PCErr message with Error-type=10 (Invalid object) and Error-value=8 (SYMBOLIC-PATH-NAME TLV missing).

The PCE MUST NOT provide a symbolic path name that conflicts with the symbolic path name of any existing LSP in the PCC. (Existing LSPs may be either statically configured, or initiated by another PCE). If there is a conflict with the symbolic path name of an existing LSP, 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 Interval timer is running (see Section 6).

If the PCC determines that the LSP parameters proposed in the PCInitiate message are unacceptable, it MUST send a PCErr message with Error-type=24 (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 send a PCErr message with Error-type=24 (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 message with Error-type=24 (PCE instantiation error) and Error-value=3 (Signaling error). The PCErr message MUST echo the SRP-ID-number of the PCInitiate message. The PCEP-ERROR object SHOULD include the RSVP_ERROR_SPEC TLV (if an RSVP ERROR_SPEC object 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 [RFC8231].

On successful completion of the LSP instantiation, the PCC MUST send a PCRpt message. The LSP object message MUST contain a non-zero

PLSP-ID that uniquely identifies the LSP within this PCC, and MUST have the Create flag (Section 5.3.1) and Delegate flag set. The SRP object MUST contain an SRP-ID-number that echoes the value from the PCInitiate message that triggered the setup. The PCRpt MUST include the attributes that the PCC used to instantiate the LSP.

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 with Error-type=19 (Invalid Operation) and Error-value=6 (PCE-initiated LSP 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 PCInitiate messages pending for a particular PCC, or on the time it waits for a response (positive or negative) to a PCInitiate message from a PCC and MAY take further action (such as closing the session or removing all its LSPs) if this limit is reached.

5.3.1. The Create Flag

The LSP object is defined in [RFC8231] and included here for easy reference with the addition of the new C flag.

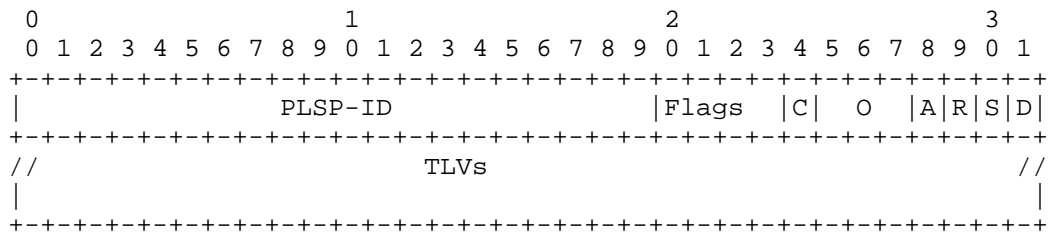


Figure 4: The LSP Object format

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).

5.3.2. The SPEAKER-ENTITY-ID TLV

The optional SPEAKER-ENTITY-ID TLV defined in [RFC8232] MAY be included in the LSP object in a PCRpt message, as an optional TLV for LSPs for which the C flag is 1. The SPEAKER-ENTITY-ID TLV identifies

the PCE which initiated the creation of the LSP on all PCEP sessions, a state that would otherwise only be known by the PCC and the PCE that initiated the LSP. If the TLV appears in a PCRpt for an LSP for which the C flag is 0, the LSP MUST be ignored and the PCE MUST send a PCErr message with Error-type=23 ("Bad parameter value") and Error-value=2 ("Speaker identity included for an LSP that is not PCE-initiated").

5.4. LSP Deletion

A PCE can initiate the removal of a PCE-initiated LSP by sending a PCInitiate 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). A PLSP-ID of zero removes all LSPs with the C flag set to 1 (in their LSP object) that are delegated to the PCE.

If the PLSP-ID is unknown, the PCC MUST send a PCErr message with Error-type=19 ("Invalid operation") and Error-value=3 ("Unknown PLSP-ID") ([RFC8231]).

If the PLSP-ID specified in the PCInitiate message is not delegated to the PCE, the PCC MUST send a PCErr message with Error-type=19 ("Invalid operation") and Error-value=1 ("LSP is not delegated") ([RFC8231]).

If the PLSP-ID specified in the PCInitiate message was not created by a PCE, the PCC MUST send a PCErr message with Error-type=19 ("Invalid operation") and Error-value=9 ("LSP is not PCE-initiated").

Following the removal of the LSP, the PCC MUST send a PCRpt as described in [RFC8231]. 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 MUST be set.

6. LSP Delegation and Cleanup

The PCC MUST delegate PCE-initiated LSPs to the PCE upon instantiation. The PCC MUST set the delegation bit to 1 in the PCRpt that includes the assigned PLSP-ID.

The PCC MUST NOT revoke the delegation for a PCE-initiated LSP on an active PCEP session. Therefore, all PCRpt messages from the PCC to the PCE that owns the delegation MUST have the delegation bit set to 1. If the PCE that owns the delegation receives a PCRpt message with the delegation bit set to 0 then it MUST send a PCErr message with Error-type=19 ("Invalid Operation") and Error-value=7 ("Delegation for PCE-initiated LSP cannot be revoked"). The PCE MAY further react by closing the session.

Control over a PCE-initiated LSP can revert to the PCC in two ways. A PCE MAY return a delegation to the PCC to allow for LSP transfer between PCEs. Alternatively, the PCC gains control of an LSP if the PCEP session that it was delegated on fails and the Redelegating Timeout Interval timer expires. In both cases, the LSP becomes an orphan until the expiration of the State Timeout Interval timer ([RFC8231]).

The PCC MAY attempt to redelegate an orphaned LSP by following the procedures of [RFC8231]. Alternatively, if the orphaned LSP was PCE-initiated, then a PCE MAY obtain control over it, as follows.

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. If the PCC receives a PCInitiate message with a PLSP-ID pointing to an orphaned PCE-initiated LSP, then it MUST redelegate that LSP to the PCE. Any other non-zero PLSP-ID MUST result in the generation of a PCErr message using the rules described in Section 5.4. The State Timeout Interval timer for the LSP 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 Interval 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 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 MUST support removal of PCE-initiated LSPs as one of the behaviors applied on expiration of the State Timeout Interval timer. The behavior MUST be picked based on local policy, and can result either in LSP removal, or in reverting to operator-defined default parameters.

7. LSP State Synchronization

LSP State Synchronization procedures are described in section 5.4 of [RFC8231]. During State Synchronization, a PCC reports the state of its LSPs to the PCE using PCRpt messages, setting the SYNC flag in the LSP Object. For PCE-initiated LSPs, the PCC MUST also set the Create Flag in the LSP Object and MAY include the SPEAKER-ENTITY-ID TLV identifying the PCE that requested the LSP creation. At the end of state synchronization, the PCE SHOULD send a PCInitiate message to initiate any missing LSPs and/or remove any LSPs that are not wanted. Under some circumstances, depending on the deployment, it might be preferable for a PCE not to send this PCInitiate immediately, or at all. For example, the PCC may be a slow device, or the operator might prefer not to disrupt active flows.

8. 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 [RFC7942]. 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 7942, "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.

9. IANA Considerations

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

9.1. PCEP Messages

IANA is requested to confirm the early allocation of the following new message type 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
12	LSP Initiate Request	This document

Note to IANA: The early allocation was done for a message called "Initiate". This name has changed to "LSP Initiate Request" as above.

9.2. LSP Object

[RFC8231] defines the LSP Object and requests that IANA creates a registry to manage the value of the LSP Object's Flag field. IANA is requested to allocate a new bit in the LSP Object Flag Field registry, as follows:

Bit	Description	Reference
4	Create	This document

9.3. SRP object

This document requests that a new sub-registry, named "SRP Object Flag Field", is created within the "Path Computation Element Protocol (PCEP) Numbers" registry to manage the Flag field of the SRP object. New values are to be assigned by Standards Action [RFC8126]. Each bit should be tracked with the following qualities: bit number (counting from bit 0 as the most significant bit), description and defining RFC.

The following values are defined in this document:

Bit	Description	Reference
31	LSP-Remove	This document

9.4. STATEFUL-PCE-CAPABILITY TLV

[RFC8231] defines the STATEFUL-PCE-CAPABILITY TLV and requests that IANA creates a registry to manage the value of the STATEFUL-PCE-CAPABILITY TLV's Flag field. IANA is requested to allocate a new bit in the STATEFUL-PCE-CAPABILITY TLV Flag Field registry, as follows:

Bit	Description	Reference
29	I (LSP-INSTITUTION-CAPABILITY)	This document

9.5. PCEP-Error Object

IANA is requested to confirm the early allocation of the following new 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
10	Invalid Object
19	Error-value=8: SYMBOLIC-PATH-NAME TLV missing Invalid operation
23	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 PCInitiate message Error-value=9: LSP is not PCE-initiated Error-value=10: PCE-initiated operation-frequency limit reached Bad parameter value
24	Error-value=1: SYMBOLIC-PATH-NAME in use Error-value=2: Speaker identity included for an LSP that is not PCE-initiated LSP instantiation error Error-value=1: Unacceptable instantiation parameters Error-value=2: Internal error Error-value=3: Signaling error

10. Security Considerations

The security considerations described in [RFC8231] apply to the extensions described in this document. Additional considerations related to a malicious PCE are introduced.

10.1. 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 with Error-type=19 ("Invalid Operation") and Error-value=6 ("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. A PCC can protect itself from such an attack by imposing a limit on the number of flaps per unit of time that it allows a PCE to generate. As soon as that limit is reached, a PCC MUST send a PCErr message with Error-type=19 ("Invalid Operation") and Error-value=10 ("PCE-initiated operation frequency reached") and is free to treat the session as having reached the limit in terms of resources allocated to honor PCE-initiated LSP requests, either permanently or for a locally-defined cool-off period.

10.2. 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.

11. Acknowledgements

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

12. References

12.1. Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <<https://www.rfc-editor.org/info/rfc2119>>.
- [RFC5440] Vasseur, JP., Ed. and JL. Le Roux, Ed., "Path Computation Element (PCE) Communication Protocol (PCEP)", RFC 5440, DOI 10.17487/RFC5440, March 2009, <<https://www.rfc-editor.org/info/rfc5440>>.
- [RFC5511] Farrel, A., "Routing Backus-Naur Form (RBNF): A Syntax Used to Form Encoding Rules in Various Routing Protocol Specifications", RFC 5511, DOI 10.17487/RFC5511, April 2009, <<https://www.rfc-editor.org/info/rfc5511>>.

- [RFC8231] Crabbe, E., Minei, I., Medved, J., and R. Varga, "Path Computation Element Communication Protocol (PCEP) Extensions for Stateful PCE", RFC 8231, DOI 10.17487/RFC8231, September 2017, <<https://www.rfc-editor.org/info/rfc8231>>.
- [RFC8232] 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", RFC 8232, DOI 10.17487/RFC8232, September 2017, <<https://www.rfc-editor.org/info/rfc8232>>.

12.2. Informative References

- [RFC4657] Ash, J., Ed. and J. Le Roux, Ed., "Path Computation Element (PCE) Communication Protocol Generic Requirements", RFC 4657, DOI 10.17487/RFC4657, September 2006, <<https://www.rfc-editor.org/info/rfc4657>>.
- [RFC7942] Sheffer, Y. and A. Farrel, "Improving Awareness of Running Code: The Implementation Status Section", BCP 205, RFC 7942, DOI 10.17487/RFC7942, July 2016, <<https://www.rfc-editor.org/info/rfc7942>>.
- [RFC8051] Zhang, X., Ed. and I. Minei, Ed., "Applicability of a Stateful Path Computation Element (PCE)", RFC 8051, DOI 10.17487/RFC8051, January 2017, <<https://www.rfc-editor.org/info/rfc8051>>.
- [RFC8126] Cotton, M., Leiba, B., and T. Narten, "Guidelines for Writing an IANA Considerations Section in RFCs", BCP 26, RFC 8126, DOI 10.17487/RFC8126, June 2017, <<https://www.rfc-editor.org/info/rfc8126>>.

Authors' Addresses

Edward Crabbe
Individual Contributor

Email: edward.crabbe@gmail.com

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

Email: inaminei@google.com

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.tech

PCE Working Group
Internet-Draft
Intended status: Standards Track
Expires: August 10, 2022

Y. Lee
Samsung
H. Zheng
Huawei Technologies
O. G. de Dios
Telefonica
Victor Lopez
Nokia
Z. Ali
Cisco Systems
February 10, 2022

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

draft-ietf-pce-pcep-stateful-pce-gmpls-17

Abstract

The Path Computation Element (PCE) facilitates Traffic Engineering (TE) based path calculation in large, multi-domain, multi-region, or multi-layer networks. The PCE communication Protocol (PCEP) has been extended to support stateful PCE functions where the PCE retains information about the paths already present in the network, but those extensions are technology-agnostic. This memo provides extensions required for PCEP so as to enable the usage of a stateful PCE capability in GMPLS-controlled networks.

Status of this Memo

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

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

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

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 August 10, 2022.

Copyright Notice

Copyright (c) 2022 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

Table of Contents	2
1. Introduction	3
2. Conventions used in this document	4
3. General Context of Stateful PCE and PCEP for GMPLS	4
4. Main Requirements	5
5. Overview of Stateful PCEP Extensions for GMPLS Networks	6
5.1. Capability Advertisement for Stateful PCEP in GMPLS	6
5.2. LSP Synchronization	6
5.3. LSP Delegation and Cleanup	7
5.4. LSP Operations	7
6. Extension of Existing PCEP Messages	7
6.1. The PCRpt Message	7
6.2. The PCUpd Message	9
6.3. The PCInitiate Message	9
7. PCEP Object Extensions	11
7.1. Existing Extensions used for Stateful GMPLS	11
7.2. New Extensions	11
7.2.1. OPEN Object Extension GMPLS-CAPABILITY TLV	11
7.2.2. New LSP Exclusion Sub-object in the XRO	12
7.2.3. SRP Extension	13

8. Update to Error Handling	13
8.1. Error Handling in LSP Re-optimization	13
8.2. Error Handling in Route Exclusion	13
8.3. Error Handling for generalized END-POINTS	14
9. Implementation	14
9.1. Huawei Technologies	14
10. IANA Considerations.....	15
10.1. New GMPLS-CAPABILITY	15
10.2. New Sub-object for the Exclude Route Object	15
10.3. Flag Field for new XRO Sub-object	15
10.4. New "B" Flag in the SRP Object	16
10.5. New PCEP Error Codes	16
11. Manageability Considerations	16
11.1. Requirements on Other Protocols	17
12. Security Considerations	17
13. Acknowledgement	17
14. References	17
14.1. Normative References	17
14.2. Informative References	18
15. Contributors' Address	19
Authors' Addresses	21

1. Introduction

[RFC4655] 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). Such a PCE is usually referred as a stateless PCE. To request path computation services to a PCE, [RFC5440] 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 [RFC5440] mainly focuses on MPLS networks and the PCEP extensions needed for GMPLS-controlled networks are provided in [RFC8779].

Stateful PCEs are shown to be helpful in many application scenarios, in both MPLS and GMPLS networks, as illustrated in [RFC8051]. Further discussion of concept of a stateful PCE can be found in [RFC7399]. In order for these applications to be able to exploit the capability of stateful PCEs, extensions to PCEP are required.

[RFC8051] describes how a stateful PCE can be applicable to solve various problems for MPLS-TE and GMPLS networks and the benefits it brings to such deployments.

[RFC8231] provides the fundamental extensions needed for stateful PCE to support general functionality. Furthermore, [RFC8281] describes the setup and teardown of PCE-initiated LSPs under the active stateful PCE model, without the need for local configuration on the PCC. However, both the documents left out the specification for technology-specific objects/TLVs, and do not cover the GMPLS networks (e.g., WSON, OTN, SONET/ SDH, etc. technologies).

This document focuses on the extensions that are necessary in order for the deployment of stateful PCEs and the requirements for remote-initiated LSPs in GMPLS-controlled networks. Section 3 provides General context of Stateful PCE and PCEP for GMPLS are provided in Section 3, and PCE initiation requirement for GMPLS is provided in section 4. Protocol extensions are included in section 5, as a solution to address such requirements.

2. Conventions used in this document

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "NOT RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in BCP 14 [RFC2119] [RFC8174] when, and only when, they appear in all capitals, as shown here.

3. General Context of Stateful PCE and PCEP for GMPLS

This section is built on the basis of Stateful PCE in [RFC8231] and PCEP for GMPLS in [RFC8779].

The operation for Stateful PCE on LSPs can be divided into two types, active stateful PCE and passive stateful PCE.

For active stateful PCE, a PCUpd message is sent from PCE to PCC to update the LSP state for the LSP delegated to the PCE. Any changes to the delegated LSPs generate a PCRpt message from the PCC to PCE to convey the changes of the LSP. Any modifications to the Objects/TLVs that are identified in this document to support GMPLS technology-specific attributes will be carried in the PCRpt and PCUpd messages.

For passive stateful PCEs, PCReq/PCRep messages are used to convey path computation instructions. GMPLS-technology specific Objects and TLVs are defined in [RFC8779], so this document just points at that work and only adds the stateful PCE aspects where applicable. Passive Stateful PCE makes use of PCRpt messages when reporting LSP State changes sent by PCC to PCEs. Any modifications to the

Objects/TLVs that are identified in this document to support GMPLS technology-specific attributes will be carried in the PCRpt message.

Furthermore, the LSP Initiation function of PCEP is defined in [RFC8281] to allow the PCE to initiate LSP establishment after the path is computed. PCInitiate messages are used to trigger the end node to set up the LSP. Any modifications to the Objects/TLVs that are identified in this document to support GMPLS technology-specific attributes will be carried in the PCInitiate messages.

[RFC8779] defines GMPLS-technology specific Objects/TLVs in stateless PCEP, and this document makes use of these Objects/TLVs without modifications where applicable. Where these Objects/TLVs require modifications to incorporate stateful PCE, they are described in this document. The remote-initiated LSP would follow the principle specified in [RFC8281], and GMPLS-specific extensions are also included in this document.

4. Main 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 [RFC8051]. 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 [RFC8231]. This document does not repeat the description of those protocol extensions. 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 requirements for GMPLS-specific stateful PCE are as follows:

- o Advertisement of the stateful PCE capability. This generic requirement is covered in Section 5.4 of [RFC8231]. The GMPLS CAPABILITY TLV in section 2.1 of [RFC8779] and its extension in this document MUST be advertised as well.
- o LSP operations, including LSP update, delegation and state synchronization/report are covered in [RFC8231]. This document provides extensions for its application in GMPLS-controlled networks.
- o All the PCEP messages need to be capable of indicating GMPLS-specific switching capabilities a per TE link basis. GMPLS LSP creation/modification/deletion requires knowledge of LSP switching capability (e.g., TDM, L2SC, OTN-TDM, LSC, etc.) and the generalized payload (G-PID) to be used according to

[RFC3471], [RFC3473]. It also requires the specification of data flow specific traffic parameters (also known as TSpec), which are technology specific. Such information would need to be included in various PCEP messages.

- o 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.
- o Stateful PCEP messages also need to indicate the protection context information for the LSP specified by GMPLS, as defined in [RFC4872], [RFC4873].

5. Overview of Stateful PCEP Extensions for GMPLS Networks

5.1. Capability Advertisement for Stateful PCEP in GMPLS

Capability Advertisement has been specified in [RFC8231], and can be achieved by using the "STATEFUL-PCE-CAPABILITY" in the PCEP TLV Type Indicators. Another GMPLS-CAPABILITY TLV in the PCEP TLV Type Indicators has been defined in [RFC8779]. According to [RFC8779], IANA created a registry to manage the value of the GMPLS-CAPABILITY TLV's Flag field. New bits, LSP-UPDATE-CAPABILITY (TBD1) and LSP-INSTITUTION-CAPABILITY (TBD2), are introduced as flags to indicate the capability for LSP update and remote LSP initiation in GMPLS networks.

5.2. LSP Synchronization

PCCs need to report the attributes of LSPs to the PCE to enable stateful operation of a GMPLS network. This process is known as LSP state synchronization. The LSP attributes including bandwidth, associated route, and protection information etc., are stored by the PCE in the LSP database (LSP-DB). Note that, as described in [RFC8231], the LSP state synchronization covers both the bulk reporting of LSPs at initialization as well the reporting of new or modified LSPs during normal operation. Incremental LSP-DB synchronization may be desired in a GMPLS-controlled network and it is specified in [RFC8232].

The END-POINTS object is extended for GMPLS in [RFC8779]. The END-POINTS object is carried in the PCRpt message as specified in

[RFC8623]. The END-POINTS object type for GMPLS is included in the PCRpt message as per the same.

The BANDWIDTH, LSPA, IRO and XRO objects are extended for GMPLS in [RFC8779]. These objects are carried in the PCRpt message as specified in [RFC8231] (as the attribute-list defined in Section 6.5 of [RFC5440] and extended by many other documents that define PCEP extensions for specific scenarios).

The SWITCH-LAYER object is defined in [RFC8282]. This object is carried in PCRpt message as specified in section 3.2 of [RFC8282].

5.3. LSP Delegation and Cleanup

LSP delegation and cleanup procedure specified in [RFC8231] are equally applicable to GMPLS LSPs and this document does not modify the associated usage.

5.4. LSP Operations

Both passive and active stateful PCE mechanisms in [RFC8231] are applicable in GMPLS-controlled networks. Remote LSP Initiation in [RFC8281] is also applicable in GMPLS-controlled networks.

6. Extension of Existing PCEP Messages

This section describes how the PCEP messages are extended by using Routing Backus-Naur Form (RBNF) [RFC5511] formats. Contents in this section are for informative purpose.

6.1. The PCRpt Message

According to [RFC8231], the PCRpt Message is used to report the current state of an LSP. This document extends the message in reporting the status of LSPs with GMPLS characteristics.

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], augmented in [RFC8779] with explicit label control (ELC) and Path Keys.

<actual-attribute-list> consists of the actual computed and signaled values of the <BANDWIDTH> and <metric-lists> objects defined in [RFC5440]. GENERALIZED-BANDWIDTH object has been defined in [RFC8779] to address the limitation of the BANDWIDTH object, with supporting the following:

- o Asymmetric bandwidth (different bandwidth in forward and reverse direction), as described in [RFC6387].
- o Technology specific GMPLS parameters (e.g., TSpec for SDH/SONET, G.709, ATM, MEF, etc.).

<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 many other documents that define PCEP extensions for specific scenarios.

The SRP object is OPTIONAL, and the usage is extended in the section 7.2.3 of this document.

6.2. The PCUpd Message

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], augmented in [RFC8779] with explicit label control (ELC) and Path Keys.

<intended-attribute-list> is the attribute-list defined in [RFC5440] and extended by many other documents that define PCEP extensions for specific scenarios.

The SRP object is OPTIONAL, and the usage is extended in the section 7.2.3 of this document.

6.3. The PCInitiate Message

According to [RFC8281], the PCInitiate Message is used allow remote LSP Initiation. This document extends the message in initiating LSPs with GMPLS characteristics. The format of a PCInitiate message is as follows:

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

Where:

<Common Header> is defined in [RFC5440].

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

[<PCE-initiated-lsp-list>]

```

    <PCE-initiated-lsp-request> ::= (<PCE-initiated-lsp-
instantiation>|

```

```
<PCE-initiated-lsp-deletion>)
```

$\langle \text{PCE-initiated-lsp-instantiation} \rangle ::= \langle \text{SRP} \rangle$

<LSP>

[<END-POINTS>]

<ERO>

[<attribute-list>]

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

<LSP>

The format of the PCInitiate message is unchanged from Section 5.1 of [RFC8281]. However, note the following:

- o The END-POINTS object was been extended by [RFC8779] to include a new object type called "Generalized Endpoint". A PCInitiate message used to trigger a GMPLS LSP instantiation MUST use that extension.
- o A PCInitiate message sent by a PCE to a PCC to trigger a GMPLS LSP instantiation MUST include the END-POINTS with Generalized Endpoint object type (even though it is marked as optional in the message definition).
- o The END-POINTS object MUST contain a "label request" TLV per [RFC8779]. The label request TLV is used to specify the switching type, encoding type and G-PID of the LSP being instantiated by the PCE.
- o If unnumbered endpoint addresses are used for the LSP being instantiated by the PCE, the unnumbered endpoint TLV [RFC8779] MUST be use to specify the unnumbered endpoint addresses.
- o The END-POINTS MAY contain other TLVs defined in [RFC8779].

7. PCEP Object Extensions

7.1. Existing Extensions used for Stateful GMPLS

Existing extensions defined in [RFC8779] can be used in the Stateful PCEP with no changes or slightly changes for GMPLS network control, including the following:

- o END-POINTS: Generalized END-POINTS was specified in [RFC8779] to include GMPLS capabilities. Stateful PCEP messages MUST include the END-POINTS with Generalized Endpoint object type, containing the "label request" TLV.
- o BANDWIDTH: Generalized BANDWIDTH was specified in [RFC8779] to represent GMPLS features, including asymmetric bandwidth and G-PID information.
- o LSPA: LSPA Extensions in Section 2.8 of [RFC8779] is applicable in Stateful PCEP for GMPLS networks.
- o IRO: IRO Extensions in Section 2.6 of [RFC8779] is applicable in Stateful PCEP for GMPLS networks.
- o XRO: XRO Extensions in Section 2.7 of [RFC8779] is applicable in Stateful PCEP for GMPLS networks. A new flag is defined in section 7.2.2 of this document.
- o ERO: The ERO was not extended in [RFC8779], and not in this document as well.
- o SWITCH-LAYER: SWITCHING-LAYER definition in Section 3.2 of [RFC8282] is applicable in Stateful PCEP messages for GMPLS networks.

7.2. New Extensions

7.2.1. OPEN Object Extension GMPLS-CAPABILITY TLV

In [RFC8779], IANA has allocated value 45 (GMPLS-CAPABILITY) from the "PCEP TLV Type Indicators" sub-registry. The TLV is extended with two flags to indicate the Stateful and remote initiate capability.

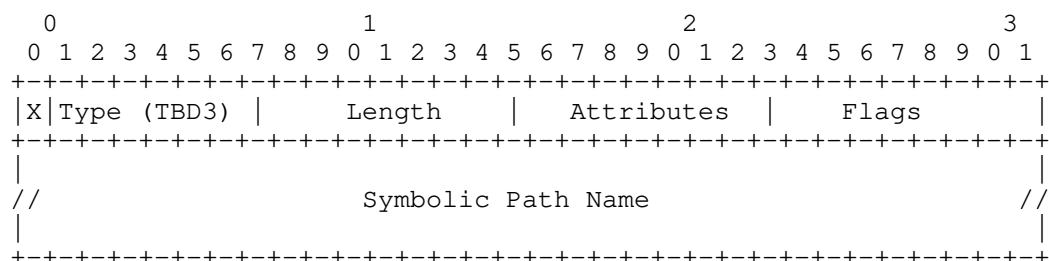
S (LSP-UPDATE-CAPABILITY(TBD1) -- 1 bit): if set to 1 by a PCC, the S flag indicates that the PCC allows modification of LSP parameters; if set to 1 by a PCE, the S flag indicates that the PCE is capable of updating LSP parameters. The LSP-UPDATE-CAPABILITY flag must be advertised by both a PCC and a PCE for PCUpd messages to be allowed on a PCEP session.

I (LSP-INSTANTIATION-CAPABILITY(TBD2) -- 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 supports instantiating LSPs. The LSP-INSTANTIATION-CAPABILITY flag must be set by both the PCC and PCE in order to enable PCE-initiated LSP instantiation.

7.2.2. New LSP Exclusion Sub-object in the XRO

[RFC5521] defines a mechanism for a PCC to request or demand that specific nodes, links, or other network resources are excluded from paths computed by a PCE. A PCC may wish to request the computation of a path that avoids all link and nodes traversed by some other LSP.

To this end this document defines a new sub-object for use with route exclusion defined in [RFC5521]. The LSP exclusion sub-object is as follows:



X bit and Attribute fields are defined in [RFC5521].

Type: Sub-object Type for an LSP exclusion sub-object. Value of TBD3. To be assigned by IANA.

Length: The Length contains the total length of the sub-object in bytes, including the Type and Length fields.

Flags: This field may be used to further specify the exclusion constraint with regard to the LSP. Currently, no values are defined.

Symbolic Path Name: This is the identifier given to an LSP and is unique in the context of the PCC address as defined in [RFC8231].

This sub-object is OPTIONAL in the exclude route object (XRO) and can be present multiple times. When a stateful PCE receives a PCReq message carrying this sub-object, it MUST search for the identified

LSP in its LSP-DB and then exclude from the new path computation all resources used by the identified LSP.

7.2.3. SRP Extension

The format of the SRP object is defined in [RFC8231]. The object is used in PCUpd and PCInitiate messages for GMPLS.

This document defines a new flag to be carried in the Flags field of the SRP object. This flag indicates a bidirectional co-routed LSP setup operation initiated by the PCE as follows:

- o B (Bidirectional LSP -- 1 bit): If set to 0, it indicates a request to create a uni-directional LSP. If set to 1, it indicates a request to create a bidirectional co-routed LSP.

The bit position is TBD4 as assigned by IANA.

8. Update to Error Handling

A PCEP-ERROR object is used to report a PCEP error and is characterized by an Error-Type that specifies the type of error and an Error-value that provides additional information about the error. In this document the following Error-Type and Error-Value are introduced.

8.1. Error Handling in LSP Re-optimization

A stateful PCE performs the re-optimization when the R bit is set in RP object. 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 = TBD5, Error value= TBD6). The PCE MAY suppress this error message on a configurable threshold.

8.2. Error Handling in Route Exclusion

This sub-object in XRO defined in section 7.2.2 of this document is OPTIONAL and can be present multiple times. When a stateful PCE receives a PCReq message carrying this sub-object, it searches for the identified LSP in its LSP-DB and then excludes from the new path computation all resources used by the identified LSP. If the stateful PCE cannot recognize one or more of the received LSP identifiers, it SHOULD send an error message PCErr reporting "The LSP state information for route exclusion purpose cannot be found"

(Error-type = TBD5, Error-value = TBD7). Optionally, it may also provide with the unrecognized identifier information to the requesting PCC using the error reporting techniques described in [RFC5440]. However, the PCE MAY suppress this error message on a configurable threshold.

8.3. Error Handling for generalized END-POINTS

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= TBD8 (label request TLV missing).

9. Implementation

[NOTE TO RFC EDITOR : This whole section and the reference to RFC 7942 is to be removed before publication as an RFC]

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 [RFC7942]. 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 [RFC7942], "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".

9.1. Huawei Technologies

- o Organization: Huawei Technologies, Co. LTD
- o Implementation: Huawei NCE-T
- o Description: PCRpt, PCUpd and PCInitiate messages for GMPLS Network

- o Maturity Level: Production
- o Coverage: Full
- o Contact: zhenghaomian@huawei.com

10. IANA Considerations

10.1. New GMPLS-CAPABILITY

[RFC8231] defines the STATEFUL-PCE-CAPABILITY TLV; per that RFC, IANA created a registry to manage the value of the STATEFUL-PCE-CAPABILITY TLV's Flag field. IANA has allocated a new bit in the STATEFUL-PCE-CAPABILITY TLV Flag Field registry, as follows:

Bit	Description	Reference
---	-----	-----
TBD1	LSP-UPDATE-CAPABILITY (S)	[This.I-D]
TBD2	LSP-INANTIATION-CAPABILITY (I)	[This.I-D]

10.2. New Sub-object for the Exclude Route Object

IANA maintains the "PCEP Parameters" registry containing a subregistry called "PCEP Objects". This registry has a subregistry for the XRO (Exclude Route Object) listing the sub-objects that can be carried in the XRO. IANA is requested to assign a further sub-object that can be carried in the XRO as follows:

Value	Description	Reference
-----+-----+-----		
TBD3	LSP Exclusion sub-object	[This.I-D]

10.3. Flag Field for new XRO Sub-object

IANA has created a registry to manage the Flag field of the LSP Exclusion sub-object in XRO object. No Flag is currently defined for this flag field in this document.

Codespace of the Flag field (LSP Exclusion sub-object)

Bit	Description	Reference
0-7	Unassigned	[This.I-D]

10.4. New "B" Flag in the SRP Object

IANA maintains a subregistry, named the "SRP Object Flag Field", within the "Path Computation Element Protocol (PCEP) Numbers" registry, to manage the Flag field of the SRP object.

IANA is requested to make an assignment from this registry as follows:

Bit ---	Description -----	Reference -----
TBD4	Bi-directional co-routed LSP	[This.I-D]

10.5. New PCEP Error Codes

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

Error Type	Meaning	Reference
TBD5	LSP state information missing	[This.I-D]
Error-value TBD6:	LSP state information unavailable for the LSP re-optimization	[This.I-D]
Error-value TBD7:	LSP state information for route exclusion purpose cannot be found	[This.I-D]

This document defines the following new Error-Value:

Error-Type	Error-Value	Reference
6	Error-value TBD8: Label Request TLV missing	[This.I-D]

11. 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 [RFC8231] should also be considered.

11.1. Requirements on Other Protocols

When the detailed route information is included for LSP state synchronization (either at the initial stage or during LSP state report process), this requires the ingress node of an LSP carry the RRO object in order to enable the collection of such information.

12. Security Considerations

This draft provides additional extensions to PCEP so as to facilitate stateful PCE usage in GMPLS-controlled networks, on top of [RFC8231]. The PCEP extensions to support GMPLS-controlled networks should be considered under the same security as for MPLS networks, as noted in [RFC7025]. Therefore, the security considerations elaborated in [RFC5440] still apply to this draft. Furthermore, [RFC8231] provides a detailed analysis of the additional security issues incurred due to the new extensions and possible solutions needed to support for the new stateful PCE capabilities and they apply to this document as well.

13. Acknowledgement

We would like to thank Adrian Farrel, Cyril Margaria, George Swallow and Jan Medved for the useful comments and discussions.

14. References

14.1. Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to indicate requirements levels", RFC 2119, March 1997.
- [RFC5440] Vasseur, J.-P., and Le Roux, JL., "Path Computation Element (PCE) Communication Protocol (PCEP)", RFC 5440, March 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.
- [RFC8174] B. Leiba, "Ambiguity of Uppercase vs Lowercase in RFC 2119 Key Words", RFC 8174, May 2017.
- [RFC8231] Crabbe, E., Medved, J., Varga, R., Minei, I., "Path Computation Element Communication Protocol (PCEP) Extensions for Stateful PCE", RFC 8231, September 2017.

- [RFC8281] Crabbe, E., Minei, I., Sivabalan, S., and R. Varga, "Path Computation Element Communication Protocol (PCEP) Extensions for PCE-Initiated LSP Setup in a Stateful PCE Model", RFC 8281, December 2017.
- [RFC8779] Margaria, C., Gonzalez de Dios, O., Zhang, F., "Path Computation Element Communication Protocol (PCEP) extensions for GMPLS", RFC 8779, July 2020.

14.2. Informative References

- [RFC7942] Sheffer, Y. and A. Farrel, "Improving Awareness of Running Code: The Implementation Status Section", BCP 205, RFC 7942, DOI 10.17487/RFC7942, July 2016, <<https://www.rfc-editor.org/info/rfc7942>>.
- [RFC8051] Zhang, X., Minei, I., et al, "Applicability of Stateful Path Computation Element (PCE) ", RFC 8051, January 2017.
- [RFC8232] 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", RFC 8232, September 2017.
- [RFC8282] Oki, E., Takeda, T., Farrel, A., and F. Zhang, "Extensions to the Path Computation Element communication Protocol (PCEP) for Inter-Layer MPLS and GMPLS Traffic Engineering", RFC 8282, December 2017.
- [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.
- [RFC4655] Farrel, A., Vasseur, J.-P., and Ash, J., "A Path Computation Element (PCE)-Based Architecture", RFC 4655, August 2006.
- [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.
- [RFC5511] Farrel, A., "Routing Backus-Naur Form (RBNF): A Syntax Used to Form Encoding Rules in Various Routing Protocol Specifications", RFC5511, April 2005.
- [RFC6387] Takacs, A., Berger, L., Caviglia, D., Fedyk, D., and J. Meuric, "GMPLS Asymmetric Bandwidth Bidirectional Label Switched Paths (LSPs)", RFC 6387, September 2011.
- [RFC7025] Otani, T., Ogaki, K., Caviglia, D., Zhang, F., and C. Margaria, "Requirements for GMPLS Applications of PCE", RFC 7025, September 2013,
- [RFC7399] Farrel, A., King, D., "Unanswered Questions in the Path Computation Element Architecture", RFC 7399, October 2014.
- [RFC8623] Palle, U., Dhody, D., Tanaka, Y., Beeram, V., "Stateful Path Computation Element (PCE) Protocol Extensions for Usage with Point-to-Multipoint TE Label Switched Paths (LSPs)" June 2019.

15. Contributors' Address

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

Dhruv Dhody
Huawei Technology
India
Email: dhruv.ietf@gmail.com

Yi Lin
Huawei Technologies
Email: yi.lin@huawei.com

Fatai Zhang
Huawei Technologies
Email: zhangfatai@huawei.com

Ramon Casellas
CTTC
Av. Carl Friedrich Gauss n7
Castelldefels, Barcelona 08860
Spain
Email: ramon.casellas@cttc.es

Siva Sivabalan
Cisco Systems
Email: msiva@cisco.com

Clarence Filsfils
Cisco Systems
Email: cfilsfil@cisco.com

Robert Varga
Pantheon Technologies
Email: nite@hq.sk

Authors' Addresses

Young Lee
Samsung
Email: younglee.tx@gmail.com

Haomian Zheng
Huawei Technologies
H1, Huawei Xiliu Beipo Village, Songshan Lake
Dongguan, Guangdong 523808
China
Email: zhenghaomian@huawei.com

Oscar Gonzalez de Dios
Telefonica
Phone: +34 913374013
Email: oscar.gonzalezdedios@telefonica.com

Victor Lopez
Nokia
Email: victor.lopez@nokia.com

Zafar Ali
Cisco Systems
Email: zali@cisco.com

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

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

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

Abstract

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

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

Status of This Memo

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

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

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

This Internet-Draft will expire on December 21, 2017.

Copyright Notice

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

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

Table of Contents

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

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

1. Introduction

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

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

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

1.1. Requirements Language

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

2. Terminology

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

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

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

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

The following terms are defined in this document:

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

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

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

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

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

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

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

Within this document, PCEP communications are described through PCC-PCE relationship. The PCE architecture also supports the PCE-PCE communication, by having the requesting PCE fill the role of a PCC, as usual.

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

3. Motivation and Objectives for Stateful PCE

3.1. Motivation

[RFC8051] presents several use cases, demonstrating scenarios that benefit from the deployment of a stateful PCE. The scenarios apply equally to MPLS-TE and GMPLS deployments.

3.1.1. Background

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

Unfortunately, this visibility in terms of global LSP state may result in a number of issues for some demand patterns, particularly within a common setup and hold priority. This issue affects online traffic engineering systems.

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

3.1.2. Why a Stateful PCE?

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

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

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

Despite these concerns in terms of implementation complexity and scalability, several TE algorithms exist today that have been demonstrated to be extremely effective in large TE systems, providing both rapid convergence and significant benefits in terms of optimality of resource usage [MXMN-TE]. All of these systems share at least two common characteristics: the requirement for both global visibility of a flow (or in this case, a TE LSP) state and for ordered control of path reservations across devices within the system being controlled. While some approaches have been suggested in order to remove the requirements for ordered control (See [MPLS-PC]), these approaches are highly dependent on traffic distribution, and do not allow for multiple simultaneous LSP priorities representing diffserv classes.

The use cases described in [RFC8051] demonstrate a need for visibility into global inter-PCC LSP state in PCE path computations, and for PCE control of sequence and timing in altering LSP path characteristics within and across PCEP sessions.

3.1.3. Protocol vs. Configuration

Note that existing configuration tools and protocols can be used to set LSP state, such as a Command Line Interface (CLI) tool. However, this solution has several shortcomings:

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

3.2. Objectives

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

- o Allow a single PCC to interact with a mix of stateless and stateful PCEs simultaneously using the same protocol, i.e. PCEP.
- o Support efficient LSP state synchronization between the PCC and one or more active or passive stateful PCEs.
- o Allow a PCC to delegate control of its LSPs to an active stateful PCE such that a given LSP is under the control of a single PCE at any given time.

- * A PCC may revoke this delegation at any time during the lifetime of the LSP. If LSP delegation is revoked while the PCEP session is up, the PCC MUST notify the PCE about the revocation.
- * A PCE may return an LSP delegation at any point during the lifetime of the PCEP session. If LSP delegation is returned by the PCE while the PCEP session is up, the PCE MUST notify the PCC about the returned delegation.
- o Allow a PCE to control computation timing and update timing across all LSPs that have been delegated to it.
- o Enable uninterrupted operation of PCC's LSPs in the event of a PCE failure or while control of LSPs is being transferred between PCEs.

4. New Functions to Support Stateful PCEs

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

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

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

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

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

LSP control delegation (C-E,E-C): a PCC grants to a PCE the right to update LSP attributes on one or more LSPs; the PCE becomes the authoritative source of the LSP's attributes as long as the delegation is in effect (See Section 5.7); the PCC may withdraw the delegation or the PCE may give up the delegation at any time.

Similarly to [RFC5440], no assumption is made about the discovery method used by a PCC to discover a set of PCEs (e.g., via static configuration or dynamic discovery) and on the algorithm used to select a PCE.

5. Overview of Protocol Extensions

5.1. LSP State Ownership

In PCEP (defined in [RFC5440]), LSP state and operation are under the control of a PCC (a PCC may be an LSR or a management station). Attributes received from a PCE are subject to PCC's local policy. The PCEP extensions described in this document do not change this behavior.

An active stateful PCE may have control of a PCC's LSPs that were delegated to it, but the LSP state ownership is retained by the PCC. In particular, in addition to specifying values for LSP's attributes, an active stateful PCE also decides when to make LSP modifications.

Retaining LSP state ownership on the PCC allows for:

- o a PCC to interact with both stateless and stateful PCEs at the same time
- o a stateful PCE to only modify a small subset of LSP parameters, i.e. to set only a small subset of the overall LSP state; other parameters may be set by the operator, for example through command line interface (CLI) commands
- o a PCC to revert delegated LSP to an operator-defined default or to delegate the LSPs to a different PCE, if the PCC get disconnected from a PCE with currently delegated LSPs

5.2. New Messages

In this document, we define the following new PCEP messages:

Path Computation State Report (PCRpt): a PCEP message sent by a PCC to a PCE to report the status of one or more LSPs. Each LSP State Report in a PCRpt message MAY contain the actual LSP's path, bandwidth, operational and administrative status, etc. An LSP Status Report carried on a PCRpt message is also used in delegation or revocation of control of an LSP to/from a PCE. The PCRpt message is described in Section 6.1.

Path Computation Update Request (PCUpd): a PCEP message sent by a PCE to a PCC to update LSP parameters, on one or more LSPs. Each LSP Update Request on a PCUpd message MUST contain all LSP parameters that a PCE wishes to be set for a given LSP. An LSP Update Request carried on a PCUpd message is also used to return LSP delegations if at any point PCE no longer desires control of an LSP. The PCUpd message is described in Section 6.2.

The new functions defined in Section 4 are mapped onto the new messages as shown in the following table.

Function	Message
Capability Advertisement (E-C,C-E)	Open
State Synchronization (C-E)	PCRpt
LSP State Report (C-E)	PCRpt
LSP Control Delegation (C-E,E-C)	PCRpt, PCUpd
LSP Update Request (E-C)	PCUpd

Table 1: New Function to Message Mapping

5.3. Error Reporting

Error reporting is done using the procedures defined in [RFC5440], and reusing the applicable error types and error values of [RFC5440] wherever appropriate. The current document defines new error values for several error types to cover failures specific to stateful PCE.

5.4. Capability Advertisement

During PCEP Initialization Phase, PCEP Speakers (PCE or PCC) advertise their support of stateful PCEP extensions. A PCEP Speaker includes the "Stateful PCE Capability" TLV, described in Section 7.1.1, in the OPEN Object to advertise its support for PCEP stateful extensions. The Stateful Capability TLV includes the 'LSP Update' Flag that indicates whether the PCEP Speaker supports LSP parameter updates.

The presence of the Stateful PCE Capability TLV in PCC's OPEN Object indicates that the PCC is willing to send LSP State Reports whenever LSP parameters or operational status changes.

The presence of the Stateful PCE Capability TLV in PCE's OPEN message indicates that the PCE is interested in receiving LSP State Reports whenever LSP parameters or operational status changes.

The PCEP extensions for stateful PCEs MUST NOT be used if one or both PCEP Speakers have not included the Stateful PCE Capability TLV in their respective OPEN message. If the PCEP Speaker on the PCC supports the extensions of this draft but did not advertise this capability, then upon receipt of PCUpd message from the PCE, it MUST generate a PCErr with error-type 19 (Invalid Operation), error-value 2 (Attempted LSP Update Request if the stateful PCE capability was not advertised)(see Section 8.5) and it SHOULD terminate the PCEP

session. If the PCEP Speaker on the PCE supports the extensions of this draft but did not advertise this capability, then upon receipt of a PCRpt message from the PCC, it MUST generate a PCErr with error-type 19 (Invalid Operation), error-value 5 (Attempted LSP State Report if stateful PCE capability was not advertised) (see Section 8.5) and it SHOULD terminate the PCEP session.

LSP delegation and LSP update operations defined in this document may only be used if both PCEP Speakers set the LSP-UPDATE-CAPABILITY Flag in the "Stateful Capability" TLV to 'Updates Allowed (U Flag = 1)'. If this is not the case and LSP delegation or LSP update operations are attempted, then a PCErr with error-type 19 (Invalid Operation) and error-value 1 (Attempted LSP Update Request for a non-delegated LSP) (see Section 8.5) MUST be generated. Note that, even if one of the PCEP speakers does not set the LSP-UPDATE-CAPABILITY flag in its "Stateful Capability" TLV, a PCE can still operate as a passive stateful PCE by accepting LSP State Reports from the PCC in order to build and maintain an up to date view of the state of the PCC's LSPs.

5.5. IGP Extensions for Stateful PCE Capabilities Advertisement

When PCCs are LSRs participating in the IGP (OSPF or IS-IS), and PCEs are either LSRs or servers also participating in the IGP, an effective mechanism for PCE discovery within an IGP routing domain consists of utilizing IGP advertisements. Extensions for the advertisement of PCE Discovery Information are defined for OSPF and for IS-IS in [RFC5088] and [RFC5089] respectively.

The PCE-CAP-FLAGS sub-TLV, defined in [RFC5089], is an optional sub-TLV used to advertise PCE capabilities. It MAY be present within the PCED sub-TLV carried by OSPF or IS-IS. [RFC5088] and [RFC5089] provide the description and processing rules for this sub-TLV when carried within OSPF and IS-IS, respectively.

The format of the PCE-CAP-FLAGS sub-TLV is included below for easy reference:

Type: 5

Length: Multiple of 4.

Value: This contains an array of units of 32 bit flags with the most significant bit as 0. Each bit represents one PCE capability.

PCE capability bits are defined in [RFC5088]. This document defines new capability bits for the stateful PCE as follows:

Bit	Capability
11	Active Stateful PCE capability
12	Passive Stateful PCE capability

Note that while active and passive stateful PCE capabilities may be advertised during discovery, PCEP Speakers that wish to use stateful PCEP MUST negotiate stateful PCEP capabilities during PCEP session setup, as specified in the current document. A PCC MAY initiate stateful PCEP capability negotiation at PCEP session setup even if it did not receive any IGP PCE capability advertisements.

5.6. State Synchronization

The purpose of State Synchronization is to provide a checkpoint-in-time state replica of a PCC's LSP state in a PCE. State Synchronization is performed immediately after the Initialization phase ([RFC5440]).

During State Synchronization, a PCC first takes a snapshot of the state of its LSPs state, then sends the snapshot to a PCE in a sequence of LSP State Reports. Each LSP State Report sent during State Synchronization has the SYNC Flag in the LSP Object set to 1. The set of LSPs for which state is synchronized with a PCE is determined by the PCC's local configuration (see more details in Section 9.1) and MAY also be determined by stateful PCEP capabilities defined in other documents, such as [I-D.ietf-pce-stateful-sync-optimizations].

The end of synchronization marker is a PCRpt message with the SYNC Flag set to 0 for an LSP Object with PLSP-ID equal to the reserved value 0 (see Section 7.3). In this case, the LSP Object SHOULD NOT include the SYMBOLIC-PATH-NAME TLV and SHOULD include the LSP-IDENTIFIERS TLV with the special value of all zeroes. The PCRpt message MUST include an empty ERO as its intended path and SHOULD NOT include the optional RRO object for its actual path. If the PCC has no state to synchronize, it SHOULD only send the end of synchronization marker.

A PCE SHOULD NOT send PCUpd messages to a PCC before State Synchronization is complete. A PCC SHOULD NOT send PCReq messages to a PCE before State Synchronization is complete. This is to allow the PCE to get the best possible view of the network before it starts computing new paths.

Either the PCE or the PCC MAY terminate the session using the PCEP session termination procedures during the synchronization phase. If the session is terminated, the PCE MUST clean up state it received from this PCC. The session reestablishment MUST be re-attempted per

the procedures defined in [RFC5440], including use of a back-off timer.

If the PCC encounters a problem which prevents it from completing the LSP state synchronization, it MUST send a PCErr message with error-type 20 (LSP State Synchronization Error) and error-value 5 (indicating an internal PCC error) to the PCE and terminate the session.

The PCE does not send positive acknowledgements for properly received synchronization messages. It MUST respond with a PCErr message with error-type 20 (LSP State Synchronization Error) and error-value 1 (indicating an error in processing the PCRpt) (see Section 8.5) if it encounters a problem with the LSP State Report it received from the PCC and it MUST terminate the session.

A PCE implementing a limit on the resources a single PCC can occupy, MUST send a PCNtf message with Notification Type 4 (Stateful PCE resource limit exceeded) and Notification Value 1 (Entering resource limit exceeded state) in response to the PCRpt message triggering this condition in the synchronization phase and MUST terminate the session.

The successful State Synchronization sequence is shown in Figure 1.

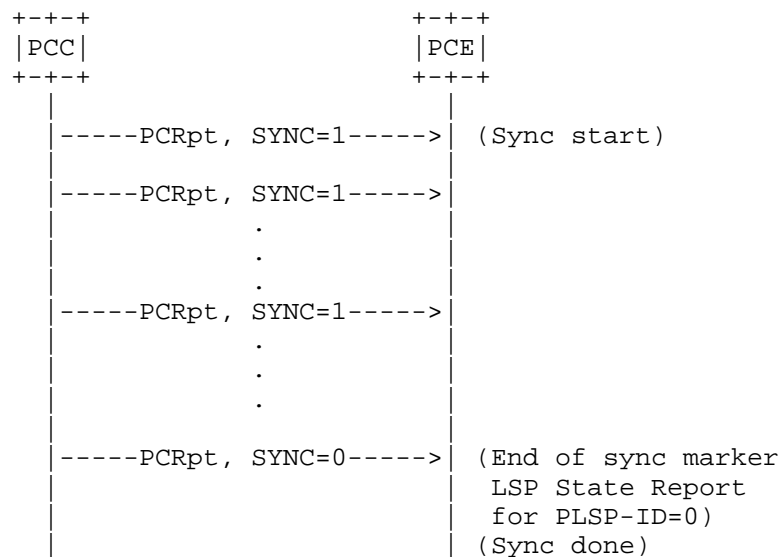


Figure 1: Successful state synchronization

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

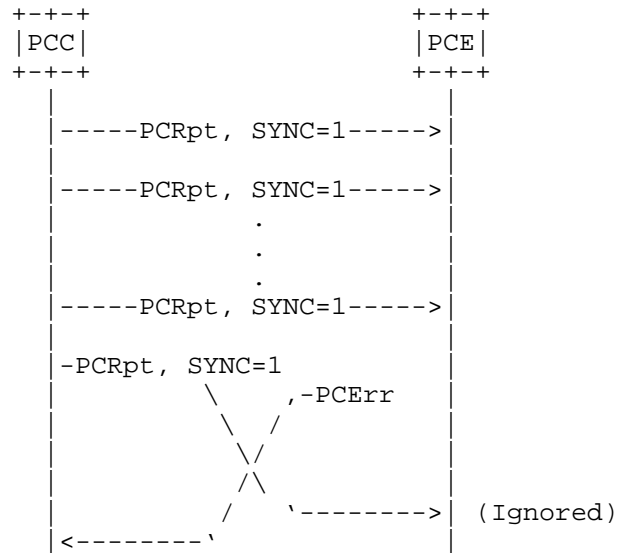


Figure 2: Failed state synchronization (PCE failure)

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

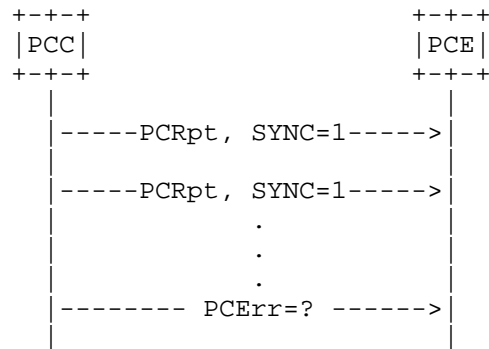


Figure 3: Failed state synchronization (PCC failure)

Optimizations to the synchronization procedures and alternate mechanisms of providing the synchronization function are outside the scope of this document and are discussed elsewhere (see [I-D.ietf-pce-stateful-sync-optimizations]).

5.7. LSP Delegation

If during Capability advertisement both the PCE and the PCC have indicated that they support LSP Update, then the PCC may choose to grant the PCE a temporary right to update (a subset of) LSP attributes on one or more LSPs. This is called "LSP Delegation", and it MAY be performed at any time after the Initialization phase, including during the State Synchronization phase.

A PCE MAY return an LSP delegation at any time if it no longer wishes to update the LSP's state. A PCC MAY revoke an LSP delegation at any time. Delegation, Revocation, and Return are done individually for each LSP.

In the event of a delegation being rejected or returned by a PCE, the PCC SHOULD react based on local policy. It can, for example, either retry delegating to the same PCE using an exponentially increasing timer or delegate to an alternate PCE.

5.7.1. Delegating an LSP

A PCC delegates an LSP to a PCE by setting the Delegate flag in LSP State Report to 1. If the PCE does not accept the LSP Delegation, it MUST immediately respond with an empty LSP Update Request which has the Delegate flag set to 0. If the PCE accepts the LSP Delegation, it MUST set the Delegate flag to 1 when it sends an LSP Update Request for the delegated LSP (note that this may occur at a later time). The PCE MAY also immediately acknowledge a delegation by sending an empty LSP Update Request which has the Delegate flag set to 1.

The delegation sequence is shown in Figure 4.

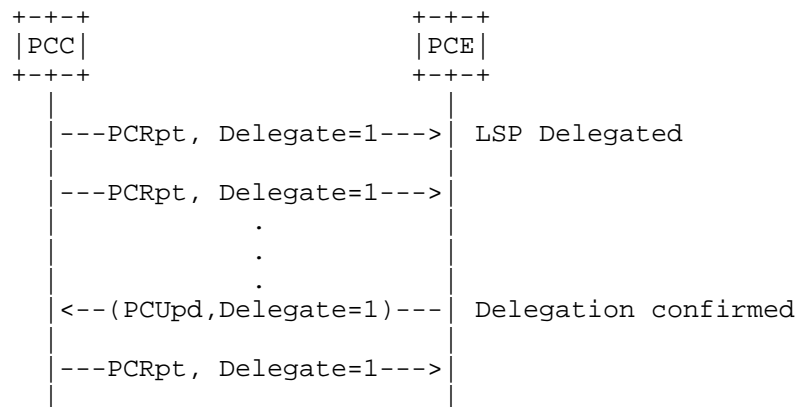


Figure 4: Delegating an LSP

Note that for an LSP to remain delegated to a PCE, the PCC MUST set the Delegate flag to 1 on each LSP State Report sent to the PCE.

5.7.2. Revoking a Delegation

5.7.2.1. Explicit Revocation

When a PCC decides that a PCE is no longer permitted to modify an LSP, it revokes that LSP's delegation to the PCE. A PCC may revoke an LSP delegation at any time during the LSP's life time. A PCC revoking an LSP delegation MAY immediately remove the updated parameters provided by the PCE and revert to the operator-defined parameters, but to avoid traffic loss, it SHOULD do so in a make-before-break fashion. If the PCC has received but not yet acted on PCUpd messages from the PCE for the LSP whose delegation is being revoked, then it SHOULD ignore these PCUpd messages when processing the message queue. All effects of all messages for which processing started before the revocation took place MUST be allowed to complete and the result MUST be given the same treatment as any LSP that had been previously delegated to the PCE (e.g. the state MAY immediately revert to the operator-defined parameters).

If a PCEP session with the PCE to which the LSP is delegated exists in the UP state during the revocation, the PCC MUST notify that PCE by sending an LSP State Report with the Delegate flag set to 0, as shown in Figure 5.

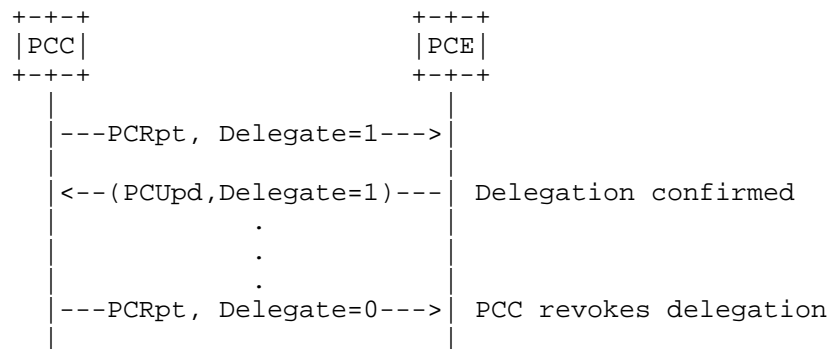


Figure 5: Revoking a Delegation

After an LSP delegation has been revoked, a PCE can no longer update LSP's parameters; an attempt to update parameters of a non-delegated LSP will result in the PCC sending a PCErr message with error-type 19 (Invalid Operation), error-value 1 (attempted LSP Update Request for a non-delegated LSP) (see Section 8.5).

5.7.2.2. Revocation on Redelegating Timeout

When a PCC's PCEP session with a PCE terminates unexpectedly, the PCC MUST wait the time interval specified in Redelegating Timeout Interval before revoking LSP delegations to that PCE and attempting to redelegate LSPs to an alternate PCE. If a PCEP session with the original PCE can be reestablished before the Redelegating Timeout Interval timer expires, LSP delegations to the PCE remain intact.

Likewise, when a PCC's PCEP session with a PCE terminates unexpectedly, and the PCC does not succeed in redelegating its LSPs, the PCC MUST wait for the State Timeout Interval before flushing any LSP state associated with that PCE. Note that the State Timeout Interval timer may expire before the PCC has redelegated the LSPs to another PCE, for example if a PCC is not connected to any active stateful PCE or if no connected active stateful PCE accepts the delegation. In this case, the PCC MUST flush any LSP state set by the PCE upon expiration of the State Timeout Interval and revert to operator-defined default parameters or behaviors. This operation SHOULD be done in a make-before-break fashion.

The State Timeout Interval MUST be greater than or equal to the Redelegating Timeout Interval and MAY be set to infinity (meaning that until the PCC specifically takes action to change the parameters set by the PCE, they will remain intact).

5.7.3. Returning a Delegation

In order to keep a delegation, a PCE MUST set the Delegate flag to 1 on each LSP Update Request sent to the PCC. A PCE that no longer wishes to update an LSP's parameters SHOULD return the LSP delegation back to the PCC by sending an empty LSP Update Request which has the Delegate flag set to 0. If a PCC receives an LSP Update Request with the Delegate flag set to 0 (whether the LSP Update Request is empty or not), it MUST treat this as a delegation return.

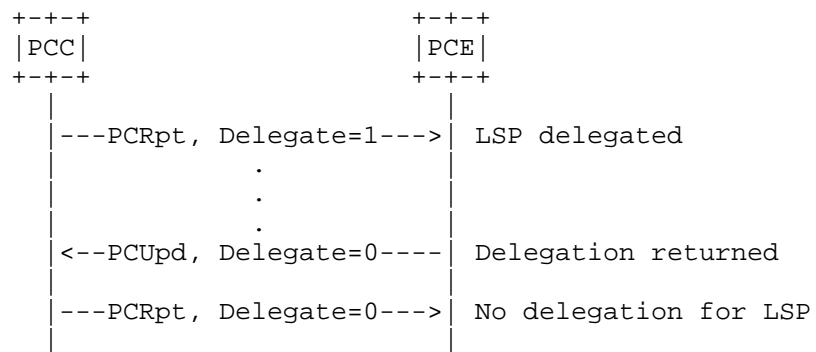


Figure 6: Returning a Delegation

If a PCC cannot delegate an LSP to a PCE (for example, if a PCC is not connected to any active stateful PCE or if no connected active stateful PCE accepts the delegation), the LSP delegation on the PCC will time out within a configurable Redelegating Timeout Interval and the PCC MUST flush any LSP state set by a PCE at the expiration of the State Timeout Interval and revert to operator-defined default parameters or behaviors.

5.7.4. Redundant Stateful PCEs

In a redundant configuration where one PCE is backing up another PCE, the backup PCE may have only a subset of the LSPs in the network delegated to it. The backup PCE does not update any LSPs that are not delegated to it. In order to allow the backup to operate in a hot-standby mode and avoid the need for state synchronization in case the primary fails, the backup receives all LSP State Reports from a PCC. When the primary PCE for a given LSP set fails, after expiry of the Redelegating Timeout Interval, the PCC SHOULD delegate to the redundant PCE all LSPs that had been previously delegated to the failed PCE. Assuming that the State Timeout Interval had been configured to be greater than the Redelegating Timeout Interval (as MANDATORY), and assuming that the primary and redundant PCEs take

similar decisions, this delegation change will not cause any changes to the LSP parameters.

5.7.5. Redelegation on PCE Failure

On failure, the goal is to: 1) avoid any traffic loss on the LSPs that were updated by the PCE that crashed 2) minimize the churn in the network in terms of ownership of the LSPs, 3) not leave any "orphan" (undelegated) LSPs and 4) be able to control when the state that was set by the PCE can be changed or purged. The values chosen for the Redelegation Timeout and State Timeout values affect the ability to accomplish these goals.

This section summarizes the behaviour with regards to LSP delegation and LSP state on a PCE failure.

If the PCE crashes but recovers within the Redelegation Timeout, both the delegation state and the LSP state are kept intact.

If the PCE crashes but does not recover within the Redelegation Timeout, the delegation state is returned to the PCC. If the PCC can redelegate the LSPs to another PCE, and that PCE accepts the delegations, there will be no change in LSP state. If the PCC cannot redelegate the LSPs to another PCE, then upon expiration of the State Timeout Interval, the state set by the PCE is removed and the LSP reverts to operator-defined parameters, which may cause a change in the LSP state. Note that an operator may choose to use an infinite State Timeout Interval if he wishes to maintain the PCE state indefinitely. Note also that flushing the state should be implemented using make-before-break to avoid traffic loss.

If there is a standby PCE, the Redelegation Timeout may be set to 0 through policy on the PCC, causing the LSPs to be redelegated immediately to the PCC, which can delegate them immediately to the standby PCE. Assuming that the PCC can redelegate the LSP to the standby PCE within the State Timeout Interval, and assuming the standby PCE takes similar decisions as the failed PCE, the LSP state will be kept intact.

5.8. LSP Operations

5.8.1. Passive Stateful PCE Path Computation Request/Response

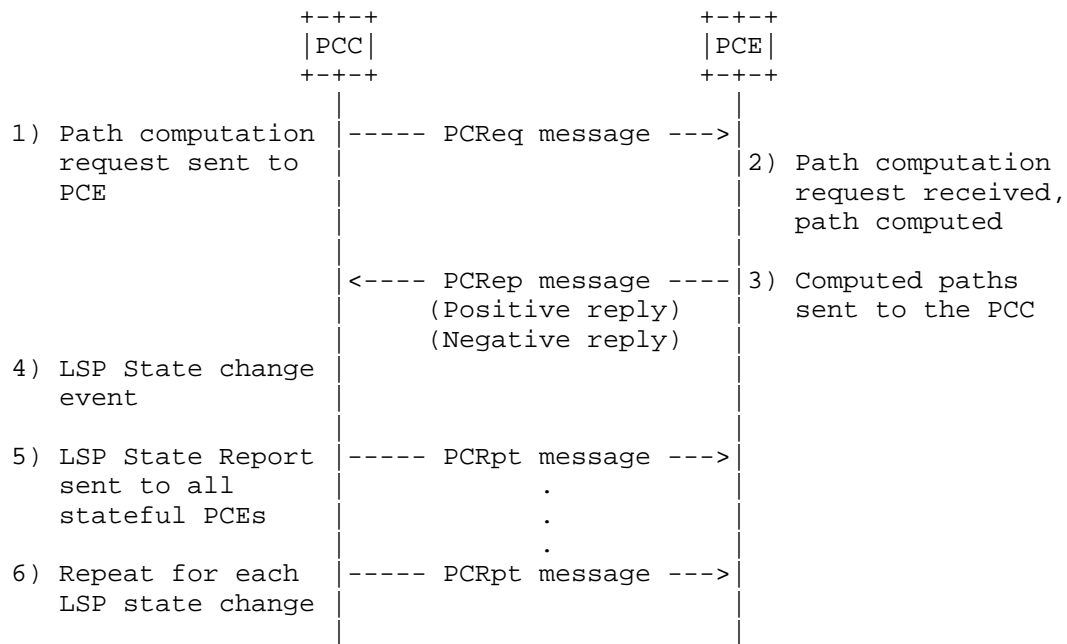


Figure 7: Passive Stateful PCE Path Computation Request/Response

Once a PCC has successfully established a PCEP session with a passive stateful PCE and the PCC's LSP state is synchronized with the PCE (i.e. the PCE knows about all PCC's existing LSPs), if an event is triggered that requires the computation of a set of paths, the PCC sends a path computation request to the PCE ([RFC5440], Section 4.2.3). The PCReq message MAY contain the LSP Object to identify the LSP for which the path computation is requested.

Upon receiving a path computation request from a PCC, the PCE triggers a path computation and returns either a positive or a negative reply to the PCC ([RFC5440], Section 4.2.4).

Upon receiving a positive path computation reply, the PCC receives a set of computed paths and starts to setup the LSPs. For each LSP, it MAY send an LSP State Report carried on a PCRpt message to the PCE, indicating that the LSP's status is "Going-up".

Once an LSP is up or active, the PCC MUST send an LSP State Report carried on a PCRpt message to the PCE, indicating that the LSP's status is 'Up' or 'Active' respectively. If the LSP could not be set up, the PCC MUST send an LSP State Report indicating that the LSP is "Down" and stating the cause of the failure. Note that due to timing constraints, the LSP status may change from 'Going-up' to 'Up' (or

'Down') before the PCC has had a chance to send an LSP State Report indicating that the status is 'Going-up'. In such cases, the PCC MAY choose to only send the PCRpt indicating the latest status ('Active', 'Up' or 'Down').

Upon receiving a negative reply from a PCE, a PCC MAY resend a modified request or take any other appropriate action. For each requested LSP, it SHOULD also send an LSP State Report carried on a PCRpt message to the PCE, indicating that the LSP's status is 'Down'.

There is no direct correlation between PCRep and PCRpt messages. For a given LSP, multiple LSP State Reports will follow a single PCRep message, as a PCC notifies a PCE of the LSP's state changes.

A PCC MUST send each LSP State Report to each stateful PCE that is connected to the PCC.

Note that a single PCRpt message MAY contain multiple LSP State Reports.

The passive stateful model for stateful PCEs is described in [RFC4655], Section 6.8.

5.8.2. Switching from Passive Stateful to Active Stateful

This section deals with the scenario of an LSP transitioning from a passive stateful to an active stateful mode of operation. When the LSP has no working path, prior to delegating the LSP, the PCC MUST first use the procedure defined in Section 5.8.1 to request the initial path from the PCE. This is required because the action of delegating the LSP to a PCE using a PCRpt message is not an explicit request to the PCE to compute a path for the LSP. The only explicit way for a PCC to request a path from PCE is to send a PCReq message. The PCRpt message MUST NOT be used by the PCC to attempt to request a path from the PCE.

When the LSP is delegated after its setup, it may be useful for the PCC to communicate to the PCE the locally configured intended configuration parameters, so that the PCE may reuse them in its computations. Such parameters MAY be acquired through an out of band channel, or MAY be communicated in the PCRpt message delegating the LSPs, by including them as part of the intended-attribute-list as explained in Section 6.1. An implementation MAY allow policies on the PCC to determine the configuration parameters to be sent to the PCE.

5.8.3. Active Stateful PCE LSP Update

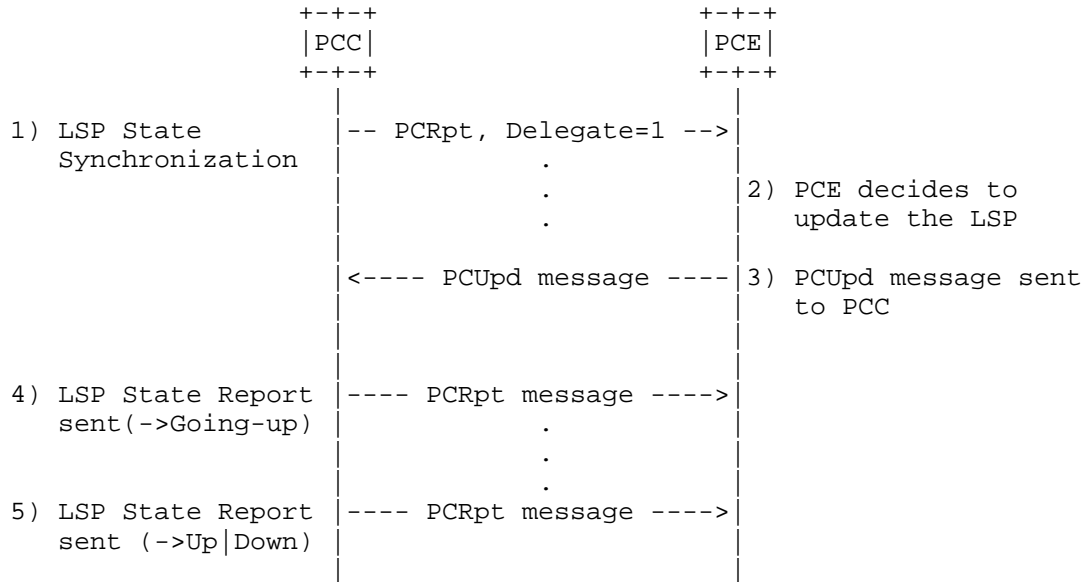


Figure 8: Active Stateful PCE

Once a PCC has successfully established a PCEP session with an active stateful PCE, the PCC's LSP state is synchronized with the PCE (i.e. the PCE knows about all PCC's existing LSPs). After LSPs have been delegated to the PCE, the PCE can modify LSP parameters of delegated LSPs.

To update an LSP, a PCE MUST send the PCC an LSP Update Request using a PCUpd message. The LSP Update Request contains a variety of objects that specify the set of constraints and attributes for the LSP's path. Each LSP Update Request MUST have a unique identifier, the SRP-ID-number, carried in the SRP (Stateful PCE Request Parameters) Object described in Section 7.2. The SRP-ID-number is used to correlate errors and state reports to LSP Update Requests. A single PCUpd message MAY contain multiple LSP Update Requests.

Upon receiving a PCUpd message the PCC starts to setup LSPs specified in LSP Update Requests carried in the message. For each LSP, it MAY send an LSP State Report carried on a PCRpt message to the PCE, indicating that the LSP's status is 'Going-up'. If the PCC decides that the LSP parameters proposed in the PCUpd message are unacceptable, it MUST report this error by including the LSP-ERROR-CODE TLV (Section 7.3.3) with LSP error-value="Unacceptable parameters" in the LSP object in the PCRpt message to the PCE. Based

on local policy, it MAY react further to this error by revoking the delegation. If the PCC receives a PCUpd message for an LSP object identified with a PLSP-ID that does not exist on the PCC, it MUST generate a PCErr with error-type 19 (Invalid Operation), error-value 3, (Attempted LSP Update Request for an LSP identified by an unknown PSP-ID) (see Section 8.5).

Once an LSP is up, the PCC MUST send an LSP State Report (PCRpt message) to the PCE, indicating that the LSP's status is 'Up'. If the LSP could not be set up, the PCC MUST send an LSP State Report indicating that the LSP is 'Down' and stating the cause of the failure. A PCC MAY compress LSP State Reports to only reflect the most up to date state, as discussed in the previous section.

A PCC MUST send each LSP State Report to each stateful PCE that is connected to the PCC.

PCErr and PCRpt messages triggered as a result of a PCUpd message MUST include the SRP-ID-number from the PCUpd. This provides correlation of requests and errors and acknowledgement of state processing. The PCC MAY compress state when processing PCUpd. In this case, receipt of a higher SRP-ID-number implicitly acknowledges processing all the updates with lower SRP-ID-number for the specific LSP (as per Section 7.2).

A PCC MUST NOT send to any PCE a Path Computation Request for a delegated LSP. Should the PCC decide it wants to issue a Path Computation Request on a delegated LSP, it MUST perform Delegation Revocation procedure first.

5.9. LSP Protection

LSP protection and interaction with stateful PCE, as well as the extensions necessary to implement this functionality will be discussed in a separate document.

5.10. PCEP Sessions

A permanent PCEP session MUST be established between a stateful PCE and the PCC. In the case of session failure, session reestablishment MUST be re-attempted per the procedures defined in [RFC5440].

6. PCEP Messages

As defined in [RFC5440], a PCEP message consists of a common header followed by a variable-length body made of a set of objects. For each PCEP message type, a set of rules is defined that specify the set of objects that the message can carry.

6.1. The PCRpt Message

A Path Computation LSP State Report message (also referred to as PCRpt message) is a PCEP message sent by a PCC to a PCE to report the current state of an LSP. A PCRpt message can carry more than one LSP State Reports. A PCC can send an LSP State Report either in response to an LSP Update Request from a PCE, or asynchronously when the state of an LSP changes. The Message-Type field of the PCEP common header for the PCRpt message is 10.

The format of the PCRpt message is as follows:

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

Where:

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

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

Where:

```
<path> ::= <intended-path>
           [<actual-attribute-list><actual-path>]
           <intended-attribute-list>
```

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

Where:

```
<intended-path> is represented by the ERO object defined in
section 7.9 of [RFC5440].
<actual-attribute-list> consists of the actual computed and
signaled values of the <BANDWIDTH> and <metric-lists> objects
defined in [RFC5440].
<actual-path> is represented by the RRO object defined in
section 7.10 of [RFC5440].
<intended-attribute-list> is the attribute-list defined in
section 6.5 of [RFC5440] and extended by PCEP extensions.
```

The SRP object (see Section 7.2) is OPTIONAL. If the PCRpt message is not in response to a PCUpd message, the SRP object MAY be omitted. When the PCC does not include the SRP object, the PCE MUST treat this as an SRP object with an SRP-ID-number equal to the reserved value 0x00000000. The reserved value 0x00000000 indicates that the state reported is not as a result of processing a PCUpd message.

If the PCRpt message is in response to a PCUpd message, the SRP object MUST be included and the value of the SRP-ID-number in the SRP Object MUST be the same as that sent in the PCUpd message that triggered the state that is reported. If the PCC compressed several PCUpd messages for the same LSP by only processing the one with the highest number, then it should use the SRP-ID-number of that request. No state compression is allowed for state reporting, e.g. PCRpt messages MUST NOT be pruned from the PCC's egress queue even if subsequent operations on the same LSP have been completed before the PCRpt message has been sent to the TCP stack. The PCC MUST explicitly report state changes (including removal) for paths it manages.

The LSP object (see Section 7.3) is REQUIRED, and it MUST be included in each LSP State Report on the PCRpt message. If the LSP object is missing, the receiving PCE MUST send a PCErr message with Error-type=6 (Mandatory Object missing) and Error-value 8 (LSP object missing).

If the LSP transitioned to non-operational state, the PCC SHOULD include the LSP-ERROR-TLV (Section 7.3.3) with the relevant LSP Error Code to report the error to the PCE.

The intended path, represented by the ERO object, is REQUIRED. If the ERO object is missing, the receiving PCE MUST send a PCErr message with Error-type=6 (Mandatory Object missing) and Error-value 9 (ERO object missing). The ERO may be empty if the PCE does not have a path for a delegated LSP.

The actual path, represented by the RRO object, SHOULD be included in PCRpt by the PCC when the path is up or active, but MAY be omitted if the path is down due to a signaling error or another failure.

The intended-attribute-list maps to the attribute-list in Section 6.5 of [RFC5440] and is used to convey the requested parameters of the LSP path. This is needed in order to support the switch from passive to active stateful PCE as described in Section 5.8.2. When included as part of the intended-attribute-list, the meaning of the BANDWIDTH object is the requested bandwidth as intended by the operator. In this case, the BANDWIDTH Object-Type of 1 SHOULD be used. Similarly, to indicate a limiting constraint, the METRIC object SHOULD be included as part of the intended-attribute-list with the B flag set and with a specific metric value. To indicate the optimization metric, the METRIC object SHOULD be included as part of the intended-attribute-list with the B flag unset and the metric value set to zero. Note that the intended-attribute-list is optional and thus may be omitted. In this case, the PCE MAY use the values in the actual-attribute-list as the requested parameters for the path.

The actual-attribute-list consists of the actual computed and signaled values of the BANDWIDTH and METRIC objects defined in [RFC5440]. When included as part of the actual-attribute-list, Object-Type 2 ([RFC5440]) SHOULD be used for the BANDWIDTH object and the C flag SHOULD be set in the METRIC object ([RFC5440]).

Note that the ordering of intended-path, actual-attribute-list, actual-path and intended-attribute-list is chosen to retain compatibility with implementations of an earlier version of this standard.

A PCE may choose to implement a limit on the resources a single PCC can occupy. If a PCRpt is received that causes the PCE to exceed this limit, the PCE MUST notify the PCC using a PCNtf message with Notification Type 4 (Stateful PCE resource limit exceeded) and Notification Value 1 (Entering resource limit exceeded state) and MUST terminate the session.

6.2. The PCUpd Message

A Path Computation LSP Update Request message (also referred to as PCUpd message) is a PCEP message sent by a PCE to a PCC to update attributes of an LSP. A PCUpd message can carry more than one LSP Update Request. The Message-Type field of the PCEP common header for the PCUpd message is 11.

The format of a PCUpd message is as follows:

```
<PCUpd Message> ::= <Common Header>
                        <update-request-list>
```

Where:

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

```
<update-request> ::= <SRP>
                      <LSP>
                      <path>
```

Where:

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

Where:

```
<intended-path> is represented by the ERO object defined in
section 7.9 of [RFC5440].
<intended-attribute-list> is the attribute-list defined in [RFC5440]
and extended by PCEP extensions.
```

There are three mandatory objects that MUST be included within each LSP Update Request in the PCUpd message: the SRP Object (see

Section 7.2), the LSP object (see Section 7.3) and the ERO object (as defined in [RFC5440], which represents the intended path. If the SRP object is missing, the receiving PCC MUST send a PCErr message with Error-type=6 (Mandatory Object missing) and Error-value=10 (SRP object missing). If the LSP object is missing, the receiving PCC MUST send a PCErr message with Error-type=6 (Mandatory Object missing) and Error-value=8 (LSP object missing). If the ERO object is missing, the receiving PCC MUST send a PCErr message with Error-type=6 (Mandatory Object missing) and Error-value=9 (ERO object missing).

The ERO in the PCUpd may be empty if the PCE cannot find a valid path for a delegated LSP. One typical situation resulting in this empty ERO carried in the PCUpd message is that a PCE can no longer find a strict SRLG-disjoint path for a delegated LSP after a link failure. The PCC SHOULD implement a local policy to decide the appropriate action to be taken: either tear down the LSP, or revoke the delegation and use a locally computed path, or keep the existing LSP.

A PCC only acts on an LSP Update Request if permitted by the local policy configured by the network manager. Each LSP Update Request that the PCC acts on results in an LSP setup operation. An LSP Update Request MUST contain all LSP parameters that a PCE wishes to be set for the LSP. A PCC MAY set missing parameters from locally configured defaults. If the LSP specified in the Update Request is already up, it will be re-signaled.

The PCC SHOULD minimize the traffic interruption, and MAY use the make-before-break procedures described in [RFC3209] in order to achieve this goal. If the make-before-break procedures are used, two paths will briefly co-exist. The PCC MUST send separate PCRpt messages for each, identified by the LSP-IDENTIFIERS TLV. When the old path is torn down after the head end switches over the traffic, this event MUST be reported by sending a PCRpt message with the LSP-IDENTIFIERS-TLV of the old path and the R bit set. The SRP-ID-number that the PCC associates with this PCRpt MUST be 0x00000000. Thus, a make-before-break operation will typically result in at least two PCRpt messages, one for the new path and one for the removal of the old path (more messages may be possible if intermediate states are reported).

If the path setup fails due to an RSVP signaling error, the error is reported to the PCE. The PCC will not attempt to resignal the path until it is prompted again by the PCE with a subsequent PCUpd message.

A PCC MUST respond with an LSP State Report to each LSP Update Request it processed to indicate the resulting state of the LSP in

the network (even if this processing did not result in changing the state of the LSP). The SRP-ID-number included in the PCRpt MUST match that in the PCUpd. A PCC MAY respond with multiple LSP State Reports to report LSP setup progress of a single LSP. In that case, the SRP-ID-number MUST be included for the first message, for subsequent messages the reserved value 0x00000000 SHOULD be used.

Note that a PCC MUST process all LSP Update Requests - for example, an LSP Update Request is sent when a PCE returns delegation or puts an LSP into non-operational state. The protocol relies on TCP for message-level flow control.

If the rate of PCUpd messages sent to a PCC for the same target LSP exceeds the rate at which the PCC can signal LSPs into the network, the PCC MAY perform state compression on its ingress queue. The compression algorithm is based on the fact that each PCUpd request contains the complete LSP state the PCE wishes to be set and works as follows: when the PCC starts processing a PCUpd message at the head of its ingress queue, it may search the queue forward for more recent PCUpd messages pertaining that particular LSP, prune all but the latest one from the queue and process only the last one as that request contains the most up-to-date desired state for the LSP. The PCC MUST NOT send PCRpt nor PCErr messages for requests which were pruned from the queue in this way. This compression step may be performed only while the LSP is not being signaled, e.g. if two PCUpd arrive for the same LSP in quick succession and the PCC started the signaling of the changes relevant to the first PCUpd, then it MUST wait until the signaling finishes (and report the new state via a PCRpt) before attempting to apply the changes indicated in the second PCUpd.

Note also that it is up to the PCE to handle inter-LSP dependencies; for example, if ordering of LSP set-ups is required, the PCE has to wait for an LSP State Report for a previous LSP before starting the update of the next LSP.

If the PCUpd cannot be satisfied (for example due to unsupported object or TLV), the PCC MUST respond with a PCErr message indicating the failure (see Section 7.3.3).

6.3. The PCErr Message

If the stateful PCE capability has been advertised on the PCEP session, the PCErr message MAY include the SRP object. If the error reported is the result of an LSP update request, then the SRP-ID-number MUST be the one from the PCUpd that triggered the error. If the error is unsolicited, the SRP object MAY be omitted. This is

equivalent to including an SRP object with SRP-ID-number equal to the reserved value 0x00000000.

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

```

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

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

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

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

<stateful-request-id-list> ::= <SRP> [<stateful-request-id-list>]

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

```

6.4. The PCReq Message

A PCC MAY include the LSP object in the PCReq message (see Section 7.3) if the stateful PCE capability has been negotiated on a PCEP session between the PCC and a PCE.

The definition of the PCReq message from [RFC5440] is extended to optionally include the LSP object after the END-POINTS object. The encoding from [RFC5440] will become:

```

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

```

Where:

```

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

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

```

6.5. The PCRep Message

A PCE MAY include the LSP object in the PCRep message (see (Section 7.3) if the stateful PCE capability has been negotiated on a PCEP session between the PCC and the PCE and the LSP object was included in the corresponding PCReq message from the PCC.

The definition of the PCRep message from [RFC5440] is extended to optionally include the LSP object after the RP object. The encoding from [RFC5440] will become:

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

Where:

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

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

7. Object Formats

The PCEP objects defined in this document are compliant with the PCEP object format defined in [RFC5440]. The P flag and the I flag of the PCEP objects defined in the current document MUST be set to 0 on transmission and SHOULD be ignored on receipt since the P and I flags are exclusively related to path computation requests.

7.1. OPEN Object

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

7.1.1. Stateful PCE Capability TLV

The STATEFUL-PCE-CAPABILITY TLV is an optional TLV for use in the OPEN Object for stateful PCE capability advertisement. Its format is shown in the following figure:

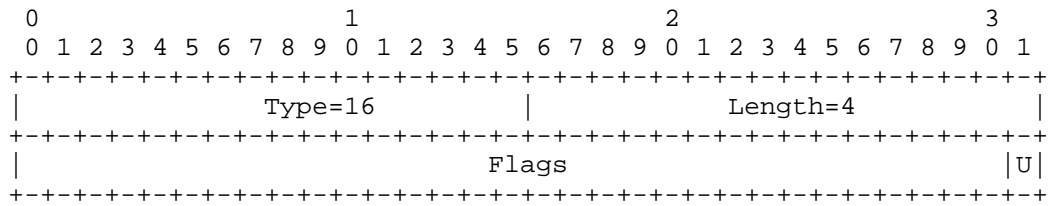


Figure 9: STATEFUL-PCE-CAPABILITY TLV format

The type (16 bits) of the TLV is 16. The length field is 16 bit-long and has a fixed value of 4.

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

U (LSP-UPDATE-CAPABILITY - 1 bit): if set to 1 by a PCC, the U Flag indicates that the PCC allows modification of LSP parameters; if set to 1 by a PCE, the U Flag indicates that the PCE is capable of updating LSP parameters. The LSP-UPDATE-CAPABILITY Flag must be advertised by both a PCC and a PCE for PCUpd messages to be allowed on a PCEP session.

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

A PCEP speaker operating in passive stateful PCE mode advertises the stateful PCE capability with the U flag set to 0. A PCEP speaker operating in active stateful PCE mode advertises the stateful PCE capability with the U Flag set to 1.

Advertisement of the stateful PCE capability implies support of LSPs that are signaled via RSVP, as well as the objects, TLVs and procedures defined in this document.

7.2. SRP Object

The SRP (Stateful PCE Request Parameters) object MUST be carried within PCUpd messages and MAY be carried within PCRpt and PCErr messages. The SRP object is used to correlate between update requests sent by the PCE and the error reports and state reports sent by the PCC.

SRP Object-Class is 33.

SRP Object-Type is 1.

The format of the SRP object body is shown in Figure 10:

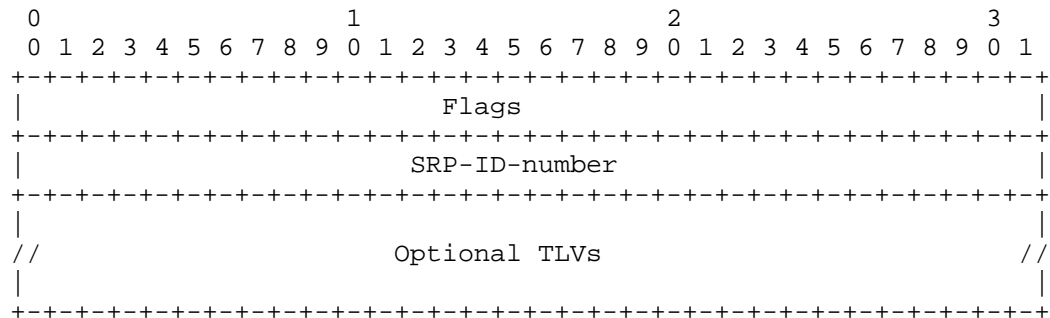


Figure 10: The SRP Object format

The SRP object body has a variable length and may contain additional TLVs.

Flags (32 bits): None defined yet.

SRP-ID-number (32 bits): The SRP-ID-number value in the scope of the current PCEP session uniquely identify the operation that the PCE has requested the PCC to perform on a given LSP. The SRP-ID-number is incremented each time a new request is sent to the PCC, and may wrap around.

The values 0x00000000 and 0xFFFFFFFF are reserved.

Optional TLVs MAY be included within the SRP object body. The specification of such TLVs is outside the scope of this document.

Every request to update an LSP receives a new SRP-ID-number. This number is unique per PCEP session and is incremented each time an operation is requested from the PCE. Thus, for a given LSP there may be more than one SRP-ID-number unacknowledged at a given time. The value of the SRP-ID-number is echoed back by the PCC in PCErr and PCRpt messages to allow for correlation between requests made by the PCE and errors or state reports generated by the PCC. If the error or report were not as a result of a PCE operation (for example in the case of a link down event), the reserved value of 0x00000000 is used for the SRP-ID-number. The absence of the SRP object is equivalent to an SRP object with the reserved value of 0x00000000. An SRP-ID-number is considered unacknowledged and cannot be reused until a PCErr or PCRpt arrives with an SRP-ID-number equal or higher for the same LSP. In case of SRP-ID-number wrapping the last SRP-ID-number before the wrapping MUST be explicitly acknowledged, to avoid a situation where SRP-ID-numbers remain unacknowledged after the wrap.

This means that the PCC may need to issue two PCUpd messages on detecting a wrap.

7.3. LSP Object

The LSP object MUST be present within PCRpt and PCUpd messages. The LSP object MAY be carried within PCReq and PCRep messages if the stateful PCE capability has been negotiated on the session. The LSP object contains a set of fields used to specify the target LSP, the operation to be performed on the LSP, and LSP Delegation. It also contains a flag indicating to a PCE that the LSP state synchronization is in progress. This document focuses on LSPs that are signaled with RSVP, many of the TLVs used with the LSP object mirror RSVP state.

LSP Object-Class is 32.

LSP Object-Type is 1.

The format of the LSP object body is shown in Figure 11:

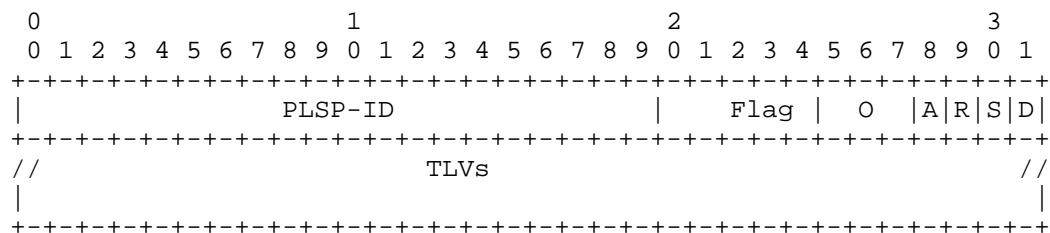


Figure 11: The LSP Object format

PLSP-ID (20 bits): A PCEP-specific identifier for the LSP. A PCC creates a unique PLSP-ID for each LSP that is constant for the lifetime of a PCEP session. The PCC will advertise the same PLSP-ID on all PCEP sessions it maintains at a given time. The mapping of the Symbolic Path Name to PLSP-ID is communicated to the PCE by sending a PCRpt message containing the SYMBOLIC-PATH-NAME TLV. All subsequent PCEP messages then address the LSP by the PLSP-ID. The values of 0 and 0xFFFFF are reserved. Note that the PLSP-ID is a value that is constant for the lifetime of the PCEP session, during which time for an RSVP-signaled LSP there might be a different RSVP identifiers (LSP-id, tunnel-id) allocated to it.

Flags (12 bits), starting from the least significant bit:

D (Delegate - 1 bit): On a PCRpt message, the D Flag set to 1 indicates that the PCC is delegating the LSP to the PCE. On a

PCUpd message, the D flag set to 1 indicates that the PCE is confirming the LSP Delegation. To keep an LSP delegated to the PCE, the PCC must set the D flag to 1 on each PCRpt message for the duration of the delegation - the first PCRpt with the D flag set to 0 revokes the delegation. To keep the delegation, the PCE must set the D flag to 1 on each PCUpd message for the duration of the delegation - the first PCUpd with the D flag set to 0 returns the delegation.

S (SYNC - 1 bit): The S Flag MUST be set to 1 on each PCRpt sent from a PCC during State Synchronization. The S Flag MUST be set to 0 in other messages sent from the PCC. When sending a PCUpd message, the PCE MUST set the S Flag to 0.

R(Remove - 1 bit): On PCRpt messages the R Flag indicates that the LSP has been removed from the PCC and the PCE SHOULD remove all state from its database. Upon receiving an LSP State Report with the R Flag set to 1 for an RSVP-signaled LSP, the PCE SHOULD remove all state for the path identified by the LSP-IDENTIFIERS TLV from its database. When the all-zeros LSP-IDENTIFIERS TLV is used, the PCE SHOULD remove all state for the PLSP-ID from its database. When sending a PCUpd message, the PCE MUST set the R Flag to 0.

A(Administrative - 1 bit): On PCRpt messages, the A Flag indicates the PCC's target operational status for this LSP. On PCUpd messages, the A Flag indicates the LSP status that the PCE desires for this LSP. In both cases, a value of '1' means that the desired operational state is active, and a value of '0' means that the desired operational state is inactive. A PCC ignores the A flag on a PCUpd message unless the operator's policy allows the PCE to control the corresponding LSP's administrative state.

O(Operational - 3 bits): On PCRpt messages, the O Field represents the operational status of the LSP.

The following values are defined:

0 - DOWN: not active.

1 - UP: signalled.

2 - ACTIVE: up and carrying traffic.

3 - GOING-DOWN: LSP is being torn down, resources are being released.

4 - GOING-UP: LSP is being signalled.

5-7 - Reserved: these values are reserved for future use.

Unassigned bits are considered reserved. They MUST be set to 0 on transmission and MUST be ignored on receipt. When sending a PCUpd message, the PCE MUST set the O Field to 0.

TLVs that may be included in the LSP Object are described in the following sections. Other optional TLVs, that are not defined in this document, MAY also be included within the LSP Object body.

7.3.1. LSP-IDENTIFIERS TLVs

The LSP-IDENTIFIERS TLV MUST be included in the LSP object in PCRpt messages for RSVP-signaled LSPs. If the TLV is missing, the PCE will generate an error with error-type 6 (mandatory object missing) and error-value 11 (LSP-IDENTIFIERS TLV missing) and close the session. The LSP-IDENTIFIERS TLV MAY be included in the LSP object in PCUpd messages for RSVP-signaled LSPs. The special value of all zeros for this TLV is used to refer to all paths pertaining to a particular PLSP-ID. There are two LSP-IDENTIFIERS TLVs, one for IPv4 and one for IPv6.

It is the responsibility of the PCC to send to the PCE the identifiers for each RSVP incarnation of the tunnel. For example, in a make-before-break scenario, the PCC MUST send a separate PCRpt for the old and for the reoptimized paths, and explicitly report removal of any of these paths using the R bit in the LSP object.

The format of the IPV4-LSP-IDENTIFIERS TLV is shown in the following figure:

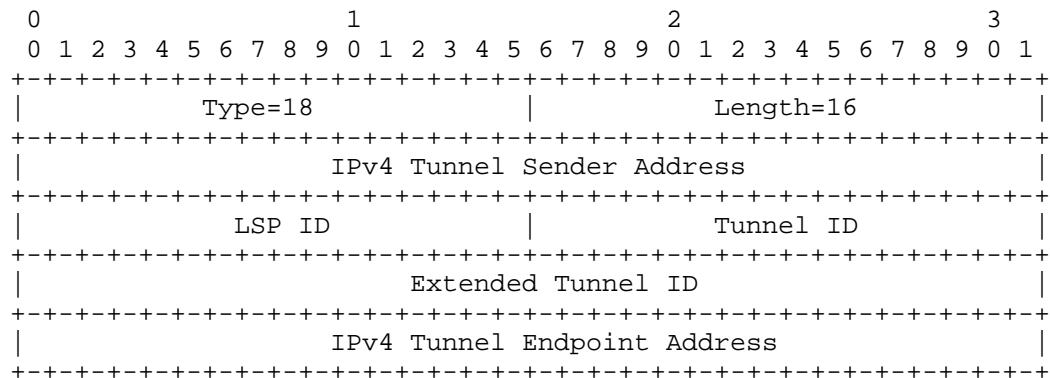


Figure 12: IPV4-LSP-IDENTIFIERS TLV format

The type (16 bits) of the TLV is 18. The length field is 16 bit-long and has a fixed value of 16. The value contains the following fields:

IPv4 Tunnel Sender Address: contains the sender node's IPv4 address, as defined in [RFC3209], Section 4.6.2.1 for the LSP_TUNNEL_IPv4 Sender Template Object.

LSP ID: contains the 16-bit 'LSP ID' identifier defined in [RFC3209], Section 4.6.2.1 for the LSP_TUNNEL_IPv4 Sender Template Object. A value of 0 MUST be used if the LSP is not yet signaled.

Tunnel ID: contains the 16-bit 'Tunnel ID' identifier defined in [RFC3209], Section 4.6.1.1 for the LSP_TUNNEL_IPv4 Session Object.

Extended Tunnel ID: contains the 32-bit 'Extended Tunnel ID' identifier defined in [RFC3209], Section 4.6.1.1 for the LSP_TUNNEL_IPv4 Session Object.

IPv4 Tunnel Endpoint Address: contains the egress node's IPv4 address, as defined in [RFC3209], Section 4.6.1.1 for the LSP_TUNNEL_IPv4 Sender Template Object.

The format of the IPV6-LSP-IDENTIFIERS TLV is shown in the following figure:

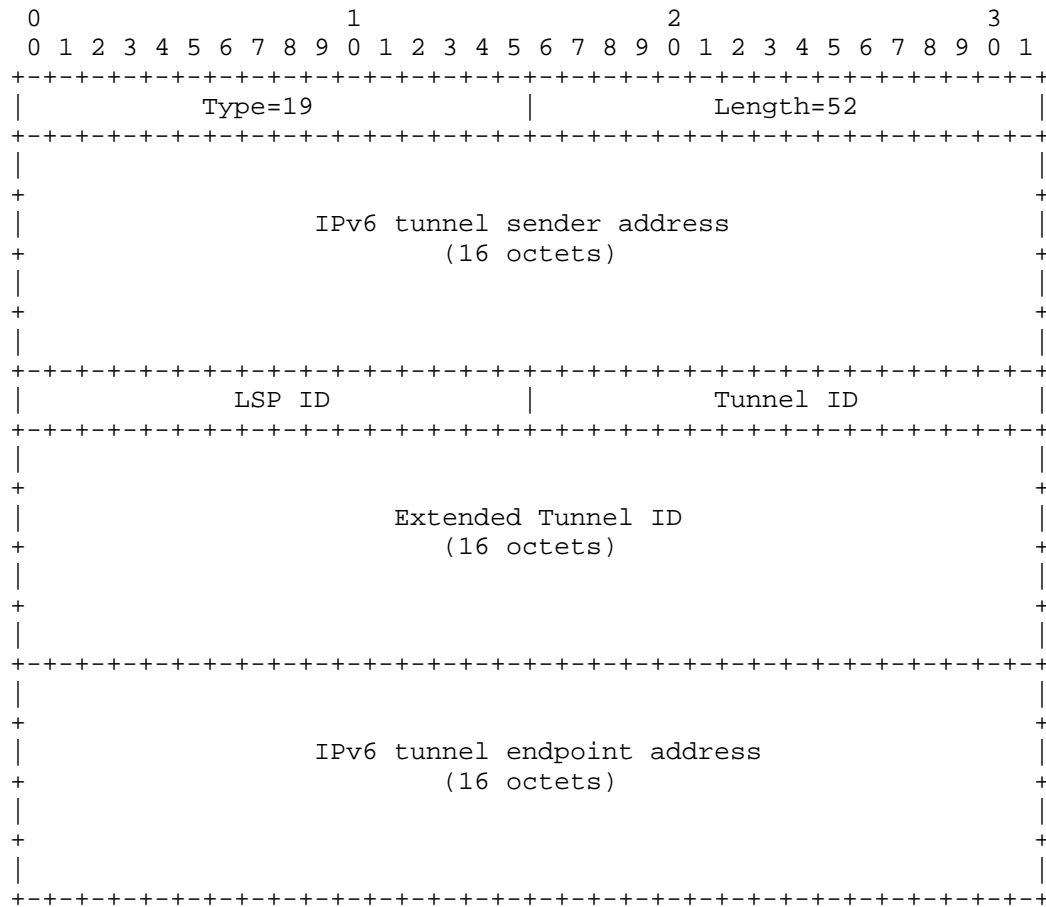


Figure 13: IPV6-LSP-IDENTIFIERS TLV format

The type (16 bits) of the TLV is 19. The length field is 16 bit-long and has a fixed value of 52. The value contains the following fields:

IPv6 Tunnel Sender Address: contains the sender node's IPv6 address, as defined in [RFC3209], Section 4.6.2.2 for the LSP_TUNNEL_IPv6 Sender Template Object.

LSP ID: contains the 16-bit 'LSP ID' identifier defined in [RFC3209], Section 4.6.2.2 for the LSP_TUNNEL_IPv6 Sender Template Object. A value of 0 MUST be used if the LSP is not yet signaled.

Tunnel ID: contains the 16-bit 'Tunnel ID' identifier defined in [RFC3209], Section 4.6.1.2 for the LSP_TUNNEL_IPv6 Session Object.

Extended Tunnel ID: contains the 128-bit 'Extended Tunnel ID' identifier defined in [RFC3209], Section 4.6.1.2 for the LSP_TUNNEL_IPv6 Session Object.

IPv6 Tunnel Endpoint Address: contains the egress node's IPv6 address, as defined in [RFC3209], Section 4.6.1.2 for the LSP_TUNNEL_IPv6 Session Object.

The Tunnel ID remains constant over the life time of a tunnel.

7.3.2. Symbolic Path Name TLV

Each LSP MUST have a symbolic path name that is unique in the PCC. The symbolic path name is a human-readable string that identifies an LSP in the network. The symbolic path name MUST remain constant throughout an LSP's lifetime, which may span across multiple consecutive PCEP sessions and/or PCC restarts. The symbolic path name MAY be specified by an operator in a PCC's configuration. If the operator does not specify a unique symbolic name for an LSP, then the PCC MUST auto-generate one.

The PCE uses the symbolic path name as a stable identifier for the LSP. If the PCEP session restarts, or the PCC restarts, or the PCC re-delegates the LSP to a different PCE, the symbolic path name for the LSP remains constant and can be used to correlate across the PCEP session instances.

The other protocol identifiers for the LSP cannot reliably be used to identify the LSP across multiple PCEP sessions, for the following reasons.

- o The PLSP-ID is unique only within the scope of a single PCEP session.
- o The LSP-IDENTIFIERS TLV is only guaranteed to be present for LSPs that are signalled with RSVP-TE, and may change during the lifetime of the LSP.

The SYMBOLIC-PATH-NAME TLV MUST be included in the LSP object in the LSP State Report (PCRpt) message when during a given PCEP session an LSP is first reported to a PCE. A PCC sends to a PCE the first LSP State Report either during State Synchronization, or when a new LSP is configured at the PCC.

The initial PCRpt creates a binding between the symbolic path name and the PLSP-ID for the LSP which lasts for the duration of the PCEP session. The PCC MAY omit the symbolic path name from subsequent LSP

State Reports for that LSP on that PCEP session, and just use the PLSP-ID.

The format of the SYMBOLIC-PATH-NAME TLV is shown in the following figure:

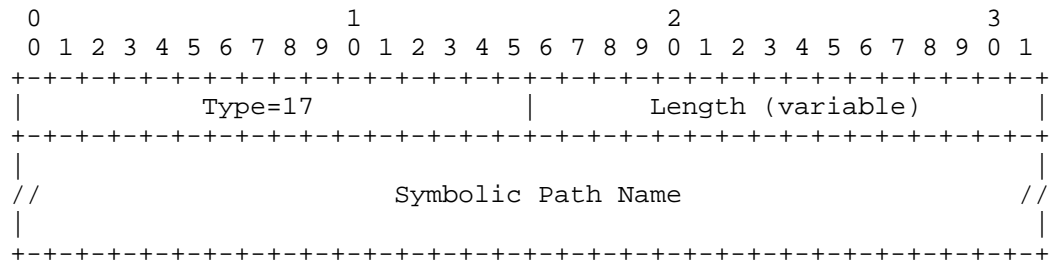


Figure 14: SYMBOLIC-PATH-NAME TLV format

```
Type (16 bits): The type is 17.
```

Length (16 bits): indicates the total length of the TLV in octets and MUST be greater than 0. The TLV MUST be zero-padded so that the TLV is 4-octet aligned.

Symbolic Path Name (variable): symbolic name for the LSP, unique in the PCC. It SHOULD be a string of printable ASCII characters, without a NULL terminator.

7.3.3. LSP Error Code TLV

The LSP Error code TLV is an optional TLV for use in the LSP object to convey error information. When an LSP Update Request fails, an LSP State Report MUST be sent to report the current state of the LSP, and SHOULD contain the LSP-ERROR-CODE TLV indicating the reason for the failure. Similarly, when a PCrpt is sent as a result of an LSP transitioning to non-operational state, the LSP-ERROR-CODE TLV SHOULD be included to indicate the reason for the transition.

The format of the LSP-ERROR-CODE TLV is shown in the following figure:

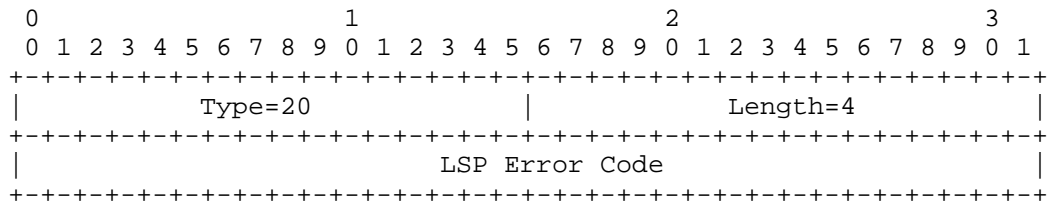


Figure 15: LSP-ERROR-CODE TLV format

The type (16 bits) of the TLV is 20. The length field is 16 bit-long and has a fixed value of 4. The value contains an error code that indicates the cause of the failure.

The following LSP Error Codes are currently defined:

Value	Meaning
1	Unknown reason
2	Limit reached for PCE-controlled LSPs
3	Too many pending LSP update requests
4	Unacceptable parameters
5	Internal error
6	LSP administratively brought down
7	LSP preempted
8	RSVP signaling error

7.3.4. RSVP Error Spec TLV

The RSVP-ERROR-SPEC TLV is an optional TLV for use in the LSP object to carry RSVP error information. It includes the RSVP ERROR_SPEC or USER_ERROR_SPEC Object ([RFC2205] and [RFC5284]) which were returned to the PCC from a downstream node. If the set up of an LSP fails at a downstream node which returned an ERROR_SPEC to the PCC, the PCC SHOULD include in the PCRpt for this LSP the LSP-ERROR-CODE TLV with LSP Error Code = "RSVP signaling error" and the RSVP-ERROR-SPEC TLV with the relevant RSVP ERROR_SPEC or USER_ERROR_SPEC Object.

The format of the RSVP-ERROR-SPEC TLV is shown in the following figure:

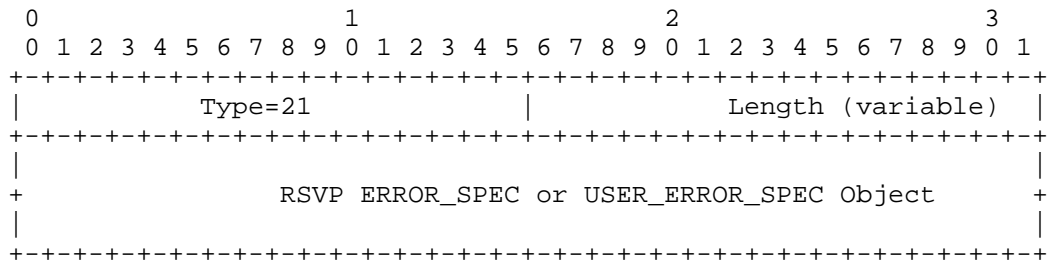


Figure 16: RSVP-ERROR-SPEC TLV format

Type (16 bits): The type is 21.

Length (16 bits): indicates the total length of the TLV in octets. The TLV MUST be zero-padded so that the TLV is 4-octet aligned.

Value (variable): contains the RSVP_ERROR_SPEC or USER_ERROR_SPEC Object: as specified in [RFC2205] and [RFC5284], including the object header.

8. IANA Considerations

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

8.1. PCE Capabilities in IGP Advertisements

IANA is requested to confirm the early allocation of the following bits in the OSPF Parameters "PCE Capability Flags" registry, and to update the reference in the registry to point to this document, when it is an RFC:

Bit	Meaning	Reference
11	Active Stateful PCE capability	This document
12	Passive Stateful PCE capability	This document

8.2. PCEP Messages

IANA is requested to confirm the early allocation of the following message types within the "PCEP Messages" sub-registry of the PCEP Numbers registry, and to update the reference in the registry to point to this document, when it is an RFC:

Value	Meaning	Reference
10	Report	This document
11	Update	This document

8.3. PCEP Objects

IANA is requested to confirm the early allocation of the following object-class values and object types within the "PCEP Objects" sub-registry of the PCEP Numbers registry, and to update the reference in the registry to point to this document, when it is an RFC:.

Object-Class Value	Name	Reference
32	LSP Object-Type 1	This document
33	SRP Object-Type 1	This document

8.4. LSP Object

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

- o Bit number (counting from bit 0 as the most significant bit)
- o Capability description
- o Defining RFC

The following values are defined in this document:

Bit	Description	Reference
0-4	Reserved	This document
5-7	Operational (3 bits)	This document
8	Administrative	This document
9	Remove	This document
10	SYNC	This document
11	Delegate	This document

8.5. PCEP-Error Object

IANA is requested to confirm the early allocation of the following Error Types and Error Values within the "PCEP-ERROR Object Error Types and Values" sub-registry of the PCEP Numbers registry, and to update the reference in the registry to point to this document, when it is an RFC:

Error-Type	Meaning
6	Mandatory Object missing
	Error-value=8: LSP Object missing
	Error-value=9: ERO Object missing
	Error-value=10: SRP Object missing
	Error-value=11: LSP-IDENTIFIERS TLV missing
19	Invalid Operation
	Error-value=1: Attempted LSP Update Request for a non-delegated LSP. The PCEP-ERROR Object is followed by the LSP Object that identifies the LSP.
	Error-value=2: Attempted LSP Update Request if the stateful PCE capability was not advertised.
	Error-value=3: Attempted LSP Update Request for an LSP identified by an unknown PLSP-ID.
	Error-value=5: Attempted LSP State Report if stateful PCE capability was not advertised.
20	LSP State synchronization error.
	Error-value=1: A PCE indicates to a PCC that it can not process (an otherwise valid) LSP State Report. The PCEP-ERROR Object is followed by the LSP Object that identifies the LSP.
	Error-value=5: A PCC indicates to a PCE that it can not complete the state synchronization,

8.6. Notification Object

IANA is requested to confirm the early allocation of the following Notification Types and Notification Values within the "Notification Object" sub-registry of the PCEP Numbers registry, and to update the reference in the registry to point to this document, when it is an RFC:

Notification-Type	Meaning
4	Stateful PCE resource limit exceeded

Notification-value=1:	Entering resource limit exceeded state
-----------------------	--

Note to IANA: the early allocation included an additional Notification value 2 for "Exiting resource limit exceeded state". This Notification value is no longer required.

8.7. PCEP TLV Type Indicators

IANA is requested to confirm the early allocation of the following TLV Type Indicator values within the "PCEP TLV Type Indicators" sub-registry of the PCEP Numbers registry, and to update the reference in the registry to point to this document, when it is an RFC:

Value	Meaning	Reference
16	STATEFUL-PCE-CAPABILITY	This document
17	SYMBOLIC-PATH-NAME	This document
18	IPV4-LSP-IDENTIFIERS	This document
19	IPV6-LSP-IDENTIFIERS	This document
20	LSP-ERROR-CODE	This document
21	RSVP-ERROR-SPEC	This document

8.8. STATEFUL-PCE-CAPABILITY TLV

This document requests that a new sub-registry, named "STATEFUL-PCE-CAPABILITY TLV Flag Field", is created within the "Path Computation Element Protocol (PCEP) Numbers" registry to manage the Flag field in the STATEFUL-PCE-CAPABILITY TLV of the PCEP OPEN object (class = 1). New values are to be assigned by Standards Action [RFC5226]. Each bit should be tracked with the following qualities:

- o Bit number (counting from bit 0 as the most significant bit)
- o Capability description
- o Defining RFC

The following values are defined in this document:

Bit	Description	Reference
31	LSP-UPDATE-CAPABILITY	This document

8.9. LSP-ERROR-CODE TLV

This document requests that a new sub-registry, named "LSP-ERROR-CODE TLV Error Code Field", is created within the "Path Computation Element Protocol (PCEP) Numbers" registry to manage the LSP Error code field of the LSP-ERROR-CODE TLV. This field specifies the reason for failure to update the LSP.

New values are to be assigned by Standards Action [RFC5226]. Each value should be tracked with the following qualities: value, description and defining RFC. The following values are defined in this document:

Value	Meaning
1	Unknown reason
2	Limit reached for PCE-controlled LSPs
3	Too many pending LSP update requests
4	Unacceptable parameters
5	Internal error
6	LSP administratively brought down
7	LSP preempted
8	RSVP signaling error

9. Manageability Considerations

All manageability requirements and considerations listed in [RFC5440] apply to PCEP extensions defined in this document. In addition, requirements and considerations listed in this section apply.

9.1. Control Function and Policy

In addition to configuring specific PCEP session parameters, as specified in [RFC5440], Section 8.1, a PCE or PCC implementation MUST allow configuring the stateful PCEP capability and the LSP Update capability. A PCC implementation SHOULD allow the operator to specify multiple candidate PCEs for and a delegation preference for each candidate PCE. A PCC SHOULD allow the operator to specify an LSP delegation policy where LSPs are delegated to the most-preferred online PCE. A PCC MAY allow the operator to specify different LSP delegation policies.

A PCC implementation which allows concurrent connections to multiple PCEs SHOULD allow the operator to group the PCEs by administrative domains and it MUST NOT advertise LSP existence and state to a PCE if the LSP is delegated to a PCE in a different group.

A PCC implementation SHOULD allow the operator to specify whether the PCC will advertise LSP existence and state for LSPs that are not

controlled by any PCE (for example, LSPs that are statically configured at the PCC).

A PCC implementation SHOULD allow the operator to specify both the Redelegating Timeout Interval and the State Timeout Interval. The default value of the Redelegating Timeout Interval SHOULD be set to 30 seconds. An operator MAY also configure a policy that will dynamically adjust the Redelegating Timeout Interval, for example setting it to zero when the PCC has an established session to a backup PCE. The default value for the State Timeout Interval SHOULD be set to 60 seconds.

After the expiration of the State Timeout Interval, the LSP reverts to operator-defined default parameters. A PCC implementation MUST allow the operator to specify the default LSP parameters. To achieve a behavior where the LSP retains the parameters set by the PCE until such time that the PCC makes a change to them, a State Timeout Interval of infinity SHOULD be used. Any changes to LSP parameters SHOULD be done in make-before-break fashion.

LSP Delegation is controlled by operator-defined policies on a PCC. LSPs are delegated individually - different LSPs may be delegated to different PCEs. An LSP is delegated to at most one PCE at any given point in time. A PCC implementation SHOULD support the delegation policy, when all PCC's LSPs are delegated to a single PCE at any given time. Conversely, the policy revoking the delegation for all PCC's LSPs SHOULD also be supported.

A PCC implementation SHOULD allow the operator to specify delegation priority for PCEs. This effectively defines the primary PCE and one or more backup PCEs to which primary PCE's LSPs can be delegated when the primary PCE fails.

Policies defined for stateful PCEs and PCCs should eventually fit in the Policy-Enabled Path Computation Framework defined in [RFC5394], and the framework should be extended to support Stateful PCEs.

9.2. Information and Data Models

The PCEP YANG module [I-D.ietf-pcep-pcep-yang] should include

- o advertised stateful capabilities and synchronization status per PCEP session
- o the delegation status of each configured LSP.

The PCEP MIB [RFC7420] could also be updated to include this information.

9.3. Liveness Detection and Monitoring

PCEP extensions defined in this document do not require any new mechanisms beyond those already defined in [RFC5440], Section 8.3.

9.4. Verifying Correct Operation

Mechanisms defined in [RFC5440], Section 8.4 also apply to PCEP extensions defined in this document. In addition to monitoring parameters defined in [RFC5440], a stateful PCC-side PCEP implementation SHOULD provide the following parameters:

- o Total number of LSP updates
- o Number of successful LSP updates
- o Number of dropped LSP updates
- o Number of LSP updates where LSP setup failed

A PCC implementation SHOULD provide a command to show for each LSP whether it is delegated, and if so, to which PCE.

A PCC implementation SHOULD allow the operator to manually revoke LSP delegation.

9.5. Requirements on Other Protocols and Functional Components

PCEP extensions defined in this document do not put new requirements on other protocols.

9.6. Impact on Network Operation

Mechanisms defined in [RFC5440], Section 8.6 also apply to PCEP extensions defined in this document.

Additionally, a PCEP implementation SHOULD allow a limit to be placed on the number of LSPs delegated to the PCE and on the rate of PCUpd and PCRpt messages sent by a PCEP speaker and processed from a peer. It SHOULD also allow sending a notification when a rate threshold is reached.

A PCC implementation SHOULD allow a limit to be placed on the rate of LSP Updates to the same LSP to avoid signaling overload discussed in Section 10.3.

10. Security Considerations

10.1. Vulnerability

This document defines extensions to PCEP to enable stateful PCEs. The nature of these extensions and the delegation of path control to PCEs results in more information being available for a hypothetical adversary and a number of additional attack surfaces which must be protected.

The security provisions described in [RFC5440] remain applicable to these extensions. However, because the protocol modifications outlined in this document allow the PCE to control path computation timing and sequence, the PCE defense mechanisms described in [RFC5440] section 7.2 are also now applicable to PCC security.

As a general precaution, it is RECOMMENDED that these PCEP extensions only be activated on authenticated and encrypted sessions across PCEs and PCCs belonging to the same administrative authority, using Transport Layer Security (TLS) [I-D.ietf-pce-pceps], as per the recommendations and best current practices in [RFC7525].

The following sections identify specific security concerns that may result from the PCEP extensions outlined in this document along with recommended mechanisms to protect PCEP infrastructure against related attacks.

10.2. LSP State Snooping

The stateful nature of this extension explicitly requires LSP status updates to be sent from PCC to PCE. While this gives the PCE the ability to provide more optimal computations to the PCC, it also provides an adversary with the opportunity to eavesdrop on decisions made by network systems external to PCE. This is especially true if the PCC delegates LSPs to multiple PCEs simultaneously.

Adversaries may gain access to this information by eavesdropping on unsecured PCEP sessions, and might then use this information in various ways to target or optimize attacks on network infrastructure. For example by flexibly countering anti-DDoS measures being taken to protect the network, or by determining choke points in the network where the greatest harm might be caused.

PCC implementations which allow concurrent connections to multiple PCEs SHOULD allow the operator to group the PCEs by administrative domains and they MUST NOT advertise LSP existence and state to a PCE if the LSP is delegated to a PCE in a different group.

10.3. Malicious PCE

The LSP delegation mechanism described in this document allows a PCC to grant effective control of an LSP to the PCE for the duration of a PCEP session. While this enables PCE control of the timing and sequence of path computations within and across PCEP sessions, it also introduces a new attack vector: an attacker may flood the PCC with PCUpd messages at a rate which exceeds either the PCC's ability to process them or the network's ability to signal the changes, either by spoofing messages or by compromising the PCE itself.

A PCC is free to revoke an LSP delegation at any time without needing any justification. A defending PCC can do this by enqueueing the appropriate PCRpt message. As soon as that message is enqueued in the session, the PCC is free to drop any incoming PCUpd messages without additional processing.

10.4. Malicious PCC

A stateful session also results in an increased attack surface by placing a requirement for the PCE to keep an LSP state replica for each PCC. It is RECOMMENDED that PCE implementations provide a limit on resources a single PCC can occupy. A PCE implementing such a limit MUST send a PCNtf message with notification-type 4 (Stateful PCE resource limit exceeded) and notification-value 1 (Entering resource limit exceeded state) upon receiving an LSP state report causing it to exceed this threshold.

Delegation of LSPs can create further strain on PCE resources and a PCE implementation MAY preemptively give back delegations if it finds itself lacking the resources needed to effectively manage the delegation. Since the delegation state is ultimately controlled by the PCC, PCE implementations SHOULD provide throttling mechanisms to prevent strain created by flaps of either a PCEP session or an LSP delegation.

11. Contributing Authors

Xian Zhang
Huawei Technology
F3-5-B R&D Center
Huawei Industrial Base, Bantian, Longgang District
Shenzhen, Guangdong 518129
P.R.China
EMail: zhang.xian@huawei.com

Dhruv Dhody
Huawei Technology

Leela Palace
Bangalore, Karnataka 560008
INDIA
EMail: dhruv.dhody@huawei.com

Siva Sivabalan
Cisco Systems, Inc.
2000 Innovation Drive
Kanata, Ontario K2K 3E8
Canada
EMail: msiva@cisco.com

12. Acknowledgements

We would like to thank Adrian Farrel, Cyril Margaria and Ramon Casellas for their contributions to this document.

We would like to thank Shane Amante, Julien Meuric, Kohei Shiimoto, Paul Schultz and Raveendra Torvi for their comments and suggestions. Thanks also to Jon Hardwick, Oscar Gonzales de Dios, Tomas Janciga, Stefan Kobza, Kexin Tang, Matej Spanik, Jon Parker, Marek Zavodsky, Ambrose Kwong, Ashwin Sampath, Calvin Ying, Mustapha Aissaoui, Stephane Litkowski and Olivier Dugeon for helpful comments and discussions.

13. References

13.1. Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <<http://www.rfc-editor.org/info/rfc2119>>.
- [RFC2205] Braden, R., Ed., Zhang, L., Berson, S., Herzog, S., and S. Jamin, "Resource ReSerVation Protocol (RSVP) -- Version 1 Functional Specification", RFC 2205, DOI 10.17487/RFC2205, September 1997, <<http://www.rfc-editor.org/info/rfc2205>>.
- [RFC3209] Awduche, D., Berger, L., Gan, D., Li, T., Srinivasan, V., and G. Swallow, "RSVP-TE: Extensions to RSVP for LSP Tunnels", RFC 3209, DOI 10.17487/RFC3209, December 2001, <<http://www.rfc-editor.org/info/rfc3209>>.
- [RFC5088] Le Roux, JL., Ed., Vasseur, JP., Ed., Ikejiri, Y., and R. Zhang, "OSPF Protocol Extensions for Path Computation Element (PCE) Discovery", RFC 5088, DOI 10.17487/RFC5088, January 2008, <<http://www.rfc-editor.org/info/rfc5088>>.

- [RFC5089] Le Roux, JL., Ed., Vasseur, JP., Ed., Ikejiri, Y., and R. Zhang, "IS-IS Protocol Extensions for Path Computation Element (PCE) Discovery", RFC 5089, DOI 10.17487/RFC5089, January 2008, <<http://www.rfc-editor.org/info/rfc5089>>.
- [RFC5284] Swallow, G. and A. Farrel, "User-Defined Errors for RSVP", RFC 5284, DOI 10.17487/RFC5284, August 2008, <<http://www.rfc-editor.org/info/rfc5284>>.
- [RFC5440] Vasseur, JP., Ed. and JL. Le Roux, Ed., "Path Computation Element (PCE) Communication Protocol (PCEP)", RFC 5440, DOI 10.17487/RFC5440, March 2009, <<http://www.rfc-editor.org/info/rfc5440>>.
- [RFC5511] Farrel, A., "Routing Backus-Naur Form (RBNF): A Syntax Used to Form Encoding Rules in Various Routing Protocol Specifications", RFC 5511, DOI 10.17487/RFC5511, April 2009, <<http://www.rfc-editor.org/info/rfc5511>>.
- [RFC8051] Zhang, X., Ed. and I. Minei, Ed., "Applicability of a Stateful Path Computation Element (PCE)", RFC 8051, DOI 10.17487/RFC8051, January 2017, <<http://www.rfc-editor.org/info/rfc8051>>.

13.2. Informative References

- [I-D.ietf-pce-gmpls-pcep-extensions]
Margarita, C., Dios, O., and F. Zhang, "PCEP extensions for GMPLS", draft-ietf-pce-gmpls-pcep-extensions-11 (work in progress), October 2015.
- [I-D.ietf-pce-pce-initiated-lsp]
Crabbe, E., Minei, I., Sivabalan, S., and R. Varga, "PCEP Extensions for PCE-initiated LSP Setup in a Stateful PCE Model", draft-ietf-pce-pce-initiated-lsp-09 (work in progress), March 2017.
- [I-D.ietf-pce-pcep-yang]
Dhody, D., Hardwick, J., Beeram, V., and j. jeffrant@gmail.com, "A YANG Data Model for Path Computation Element Communications Protocol (PCEP)", draft-ietf-pce-pcep-yang-02 (work in progress), March 2017.
- [I-D.ietf-pce-pceps]
Lopez, D., Dios, O., Wu, Q., and D. Dhody, "Secure Transport for PCEP", draft-ietf-pce-pceps-14 (work in progress), May 2017.

- [I-D.ietf-pce-stateful-sync-optimizations]
Crabbe, E., Minei, I., Medved, J., Varga, R., Zhang, X.,
and D. Dhody, "Optimizations of Label Switched Path State
Synchronization Procedures for a Stateful PCE", draft-
ietf-pce-stateful-sync-optimizations-10 (work in
progress), March 2017.
- [MPLS-PC] Chaieb, I., Le Roux, J.L., and B. Cousin, "Improved MPLS-TE
LSP Path Computation using Preemption", Global
Information Infrastructure Symposium, July 2007.
- [MXMN-TE] Danna, E., Mandal, S., and A. Singh, "Practical linear
programming algorithm for balancing the max-min fairness
and throughput objectives in traffic engineering",
INFOCOM, 2012 Proceedings IEEE Page(s): 846-854, 2012.
- [RFC2702] Awduche, D., Malcolm, J., Agogbua, J., O'Dell, M., and J.
McManus, "Requirements for Traffic Engineering Over MPLS",
RFC 2702, DOI 10.17487/RFC2702, September 1999,
<<http://www.rfc-editor.org/info/rfc2702>>.
- [RFC3031] Rosen, E., Viswanathan, A., and R. Callon, "Multiprotocol
Label Switching Architecture", RFC 3031,
DOI 10.17487/RFC3031, January 2001,
<<http://www.rfc-editor.org/info/rfc3031>>.
- [RFC3346] Boyle, J., Gill, V., Hannan, A., Cooper, D., Awduche, D.,
Christian, B., and W. Lai, "Applicability Statement for
Traffic Engineering with MPLS", RFC 3346,
DOI 10.17487/RFC3346, August 2002,
<<http://www.rfc-editor.org/info/rfc3346>>.
- [RFC3630] Katz, D., Kompella, K., and D. Yeung, "Traffic Engineering
(TE) Extensions to OSPF Version 2", RFC 3630,
DOI 10.17487/RFC3630, September 2003,
<<http://www.rfc-editor.org/info/rfc3630>>.
- [RFC4655] Farrel, A., Vasseur, J., and J. Ash, "A Path Computation
Element (PCE)-Based Architecture", RFC 4655,
DOI 10.17487/RFC4655, August 2006,
<<http://www.rfc-editor.org/info/rfc4655>>.
- [RFC4657] Ash, J., Ed. and J. Le Roux, Ed., "Path Computation
Element (PCE) Communication Protocol Generic
Requirements", RFC 4657, DOI 10.17487/RFC4657, September
2006, <<http://www.rfc-editor.org/info/rfc4657>>.

- [RFC5226] Narten, T. and H. Alvestrand, "Guidelines for Writing an IANA Considerations Section in RFCs", BCP 26, RFC 5226, DOI 10.17487/RFC5226, May 2008, <<http://www.rfc-editor.org/info/rfc5226>>.
- [RFC5305] Li, T. and H. Smit, "IS-IS Extensions for Traffic Engineering", RFC 5305, DOI 10.17487/RFC5305, October 2008, <<http://www.rfc-editor.org/info/rfc5305>>.
- [RFC5394] Bryskin, I., Papadimitriou, D., Berger, L., and J. Ash, "Policy-Enabled Path Computation Framework", RFC 5394, DOI 10.17487/RFC5394, December 2008, <<http://www.rfc-editor.org/info/rfc5394>>.
- [RFC7420] Koushik, A., Stephan, E., Zhao, Q., King, D., and J. Hardwick, "Path Computation Element Communication Protocol (PCEP) Management Information Base (MIB) Module", RFC 7420, DOI 10.17487/RFC7420, December 2014, <<http://www.rfc-editor.org/info/rfc7420>>.
- [RFC7525] Sheffer, Y., Holz, R., and P. Saint-Andre, "Recommendations for Secure Use of Transport Layer Security (TLS) and Datagram Transport Layer Security (DTLS)", BCP 195, RFC 7525, DOI 10.17487/RFC7525, May 2015, <<http://www.rfc-editor.org/info/rfc7525>>.

Authors' Addresses

Edward Crabbe
Oracle
1501 4th Ave, suite 1800
Seattle, WA 98101
US

Email: edward.crabbe@oracle.com

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.

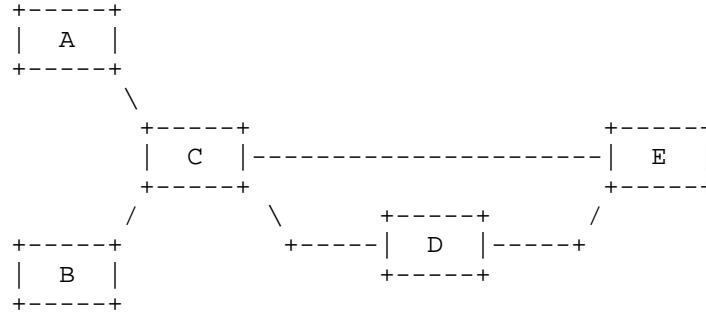


Figure 1: Reference topology 1

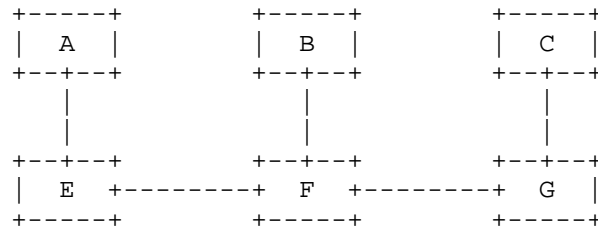


Figure 2: Reference topology 2

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.

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: Standards Track
Expires: September 28, 2017

E. Crabbe
Oracle
I. Minei
Google, Inc.
J. Medved
Cisco Systems, Inc.
R. Varga
Pantheon Technologies SRO
X. Zhang
D. Dhody
Huawei Technologies
March 27, 2017

Optimizations of Label Switched Path State Synchronization Procedures
for a Stateful PCE
draft-ietf-pce-stateful-sync-optimizations-10

Abstract

A stateful Path Computation Element (PCE) has access to not only the information disseminated by the network's Interior Gateway Protocol (IGP), but also the set of active paths and their reserved resources for its computation. The additional Label Switched Path (LSP) state information allows the PCE to compute constrained paths while considering individual LSPs and their interactions. This requires a state synchronization mechanism 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 document presents motivations for optimizations to the base state synchronization procedure and specifies the required Path Computation Element Communication Protocol (PCEP) extensions.

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 September 28, 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
2. Terminology	4
3. State Synchronization Avoidance	4
3.1. Motivation	4
3.2. State Synchronization Avoidance Procedure	4
3.2.1. IP Address change during session re-establishment	9
3.3. PCEP Extensions	10
3.3.1. LSP State Database Version Number TLV	10
3.3.2. Speaker Entity Identifier TLV	11
4. Incremental State Synchronization	12
4.1. Motivation	12
4.2. Incremental Synchronization Procedure	13
5. PCE-triggered Initial Synchronization	16
5.1. Motivation	16
5.2. PCE-triggered Initial State Synchronization Procedure	17
6. PCE-triggered Re-synchronization	18
6.1. Motivation	18
6.2. PCE-triggered State Re-synchronization Procedure	18
7. Advertising Support of Synchronization Optimizations	19
8. IANA Considerations	20
8.1. PCEP-Error Object	20

8.2. PCEP TLV Type Indicators	21
8.3. STATEFUL-PCE-CAPABILITY TLV	21
9. Manageability Considerations	21
9.1. Control of Function and Policy	21
9.2. Information and Data Models	21
9.3. Liveness Detection and Monitoring	22
9.4. Verify Correct Operations	22
9.5. Requirements On Other Protocols	22
9.6. Impact On Network Operations	22
10. Security Considerations	22
11. Acknowledgments	23
12. Contributors	23
13. References	23
13.1. Normative References	23
13.2. Informative References	24
Authors' Addresses	24

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 Interior Gateway Protocol (IGP), but also the set of active paths and their reserved resources for its computations. The additional state allows the PCE to compute constrained paths while considering individual LSPs and their interactions. This requires a state synchronization mechanism 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 document specifies following optimizations for state synchronization and the corresponding PCEP procedures and extensions:

- o State Synchronization Avoidance: To skip state synchronization if the state has survived and not changed during session restart. (See Section 3.)
- o Incremental State Synchronization: To do incremental (delta) state synchronization when possible. (See Section 4.)
- o PCE-triggered Initial Synchronization: To let PCE control the timing of the initial state synchronization. (See Section 5.)
- o PCE-triggered Re-synchronization: To let PCE re-synchronize the state for sanity check. (See Section 6.)

Support for each of the synchronization optimization capabilities is advertised during the PCEP initialization phase. See Section 7 for the new flags defined in this document. The handling of each flag is described in the relevant section.

2. Terminology

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

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

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

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

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 stateful 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 exchanges and the time it takes for the stateful PCE to become fully operational.

3.2. State Synchronization Avoidance Procedure

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 INCLUDE-DB-VERSION (S) bit in the stateful PCE capability TLV (Section 7) is advertised on a PCEP session during session startup to indicate that the LSP State Database Version Number is to be included when the LSPs are reported to the PCE. The LSP State Database Version Number, carried in LSP-DB-VERSION TLV (see Section 3.3.1), is owned by a PCC and it MUST be incremented by 1 for each successive change in the PCC's LSP state database. The LSP State Database Version Number MUST start at 1 and may wrap around. Values 0 and 0xFFFFFFFFFFFFFFFF are reserved. If either of the two values are used during LSP state (re)-synchronization, the PCE speaker receiving this value MUST send back a PCErr with Error-type 20 Error-value TBD6 (suggested value - 6) 'Received an invalid LSP DB Version Number', and close the PCEP session. 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 setup of an LSP or change in any of the LSP attributes that would trigger a report to the PCE.

If the include LSP DB version capability 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).

If both PCEP speakers set the S flag in the OPEN object's STATEFUL-PCE-CAPABILITY TLV to 1, the PCC MUST include the LSP-DB-VERSION TLV in each LSP object of the PCRpt message. If the LSP-DB-VERSION TLV is missing in a PCRpt message, the PCE will generate an error with Error-Type 6 (mandatory object missing) and Error-Value TBD1 (suggested value - 12) 'LSP-DB-VERSION TLV missing' and close the session. If the include LSP DB version capability 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 MUST ignore it were it to receive one.

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 the 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. If a PCEP speaker's LSP state database did not survive the restart of a PCEP session or at startup when the database is empty, 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 and the PCE does not wait for the end of synchronization marker [I-D.ietf-pce-stateful-pce]. Otherwise, the PCC MUST perform full state synchronization (see [I-D.ietf-pce-stateful-pce]) or incremental state synchronization (see Section 4 if this capability is advertised) to the stateful PCE. In other words, if the incremental state synchronization capability is not advertised by the peers, based on the LSP database version number match either the state synchronization is skipped or a full state synchronization is performed. If the PCC attempts to skip state synchronization, by setting the SYNC Flag to 0 and PLSP-ID to a non-zero value on the first LSP State Report from the PCC as per [I-D.ietf-pce-stateful-pce], the PCE MUST send back a PCerr with Error-Type 20 Error-Value TBD2 (suggested 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 Path Computation State Report (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 in this PCRpt message. 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.

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.

Figure 1 shows an example sequence where the state synchronization is skipped.

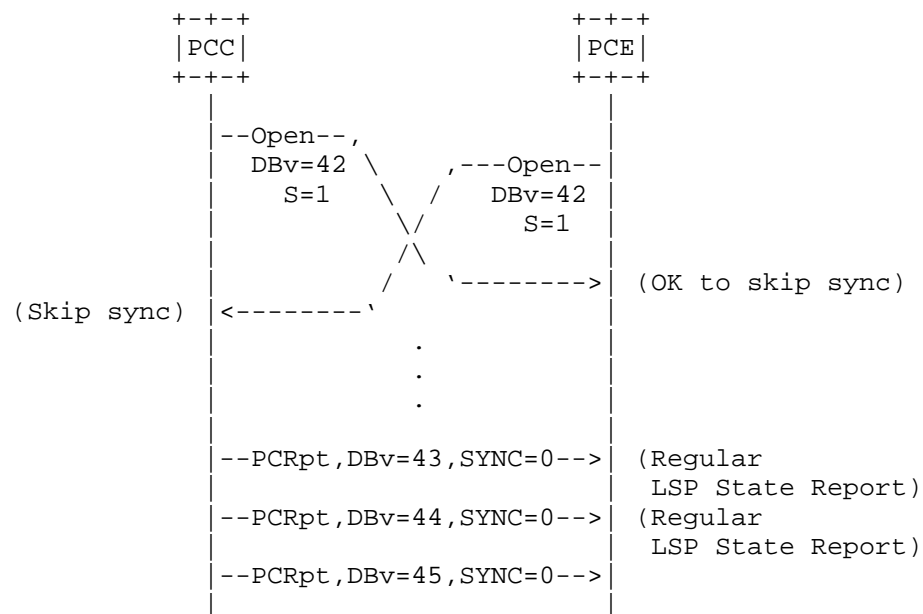


Figure 1: State Synchronization Skipped

Figure 2 shows an example sequence where the 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.

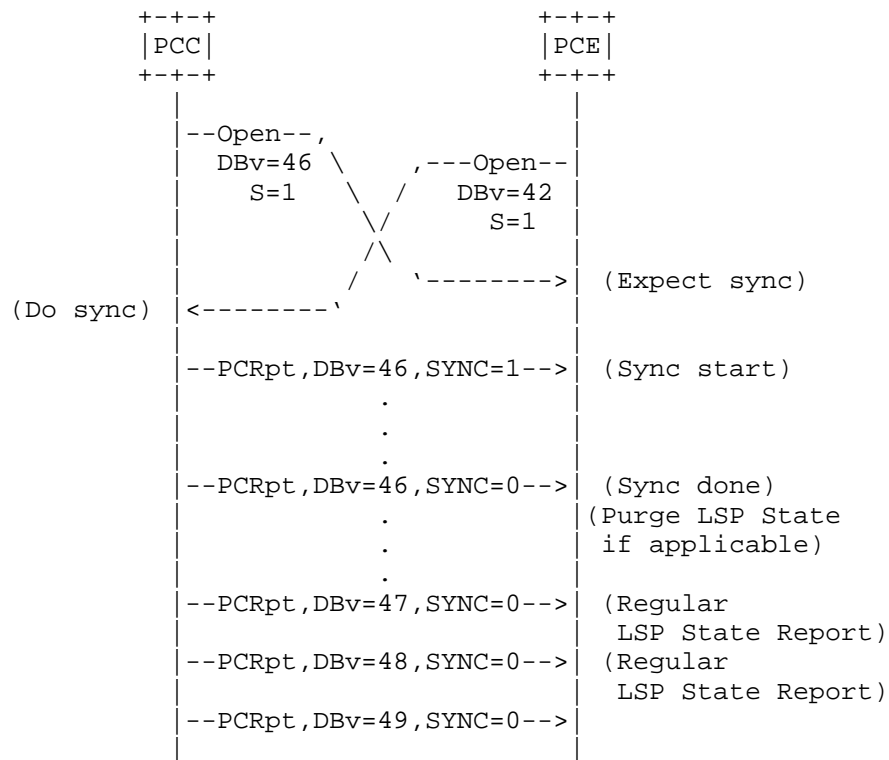


Figure 2: State Synchronization Performed

Figure 3 shows an example sequence where the state synchronization is skipped, but because one or both PCEP speakers set the S Flag to 0, the PCC does not send LSP-DB-VERSION TLVs in subsequent PCRpt messages to the PCE. If the current PCEP session restarts, the PCEP speakers will have to perform state synchronization, since the PCE does not know the PCC's latest LSP State Database Version Number information.

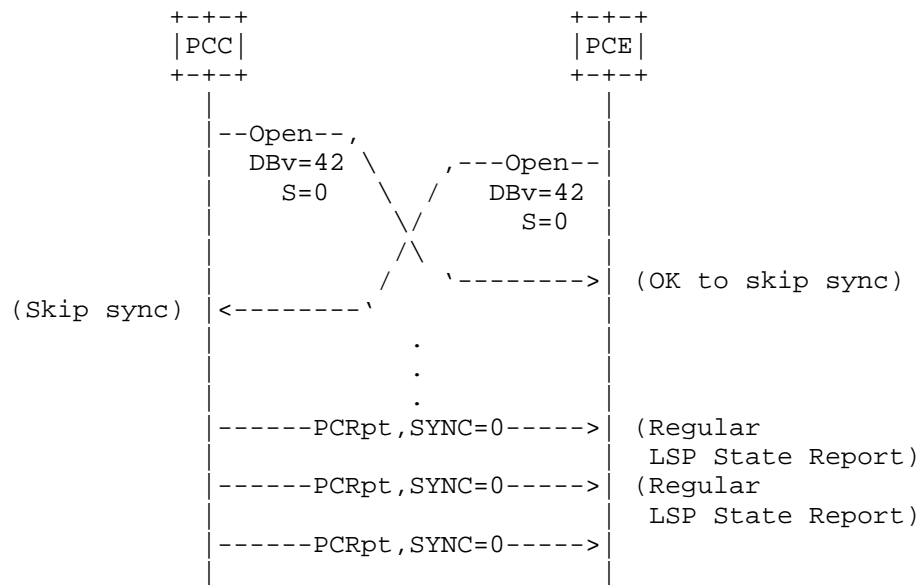


Figure 3: State Synchronization Skipped, no LSP-DB-VERSION TLVs sent from PCC

3.2.1. IP Address change during session re-establishment

There could be a case during PCEP session re-establishment when the PCC's or PCE's IP address can change. This includes, but is not limited to, the following cases:

- o A PCC could use a physical interface IP address to connect to the PCE. In this case, if the line card that the PCC connects from changes, then the PCEP session goes down and comes back up again, with a different IP address associated with a new line card.
- o The PCC or PCE may move in the network, either physically or logically, which may cause its IP address to change. For example, the PCE may be deployed as a virtual network function (VNF) and another virtualized instance of the PCE may be populated with the original PCE instance's state, but be given a different IP address.

To ensure that a PCEP peer can recognize a previously connected peer, each PCEP peer includes the SPEAKER-ENTITY-ID TLV described in Section 3.3.2, in the OPEN message.

This TLV is used during the state synchronization procedure to identify the PCEP session as a re-establishment of a previous session that went down. Then state synchronization optimizations such as state sync avoidance can be applied to this session. Note that this usage is only applicable within the State Timeout Interval [I-D.ietf-pce-stateful-pce]. After the State Timeout Interval expires, all state associated with the PCEP session is removed, which includes the SPEAKER-ENTITY-ID received. Note that the PCEP session initialization [RFC5440] procedure remains unchanged.

3.3. PCEP Extensions

A new INCLUDE-DB-VERSION (S) bit is added in the stateful capabilities TLV (see Section 7 for details).

3.3.1. 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.

This TLV is included in the LSP object in the PCRpt message to indicate the LSP DB version at the PCC. This TLV SHOULD NOT be included in other PCEP messages (PCUpd, PcReq, PCRep) and MUST be ignored if received.

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

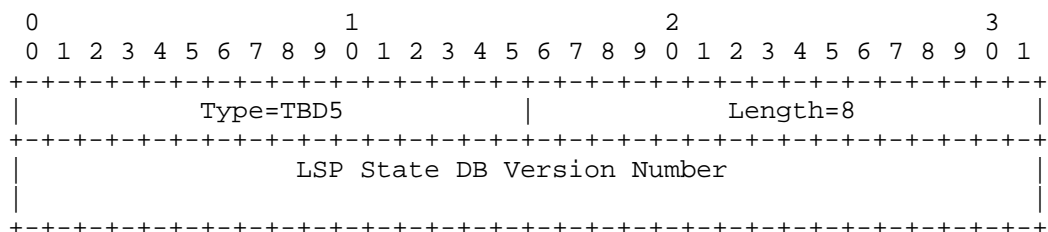
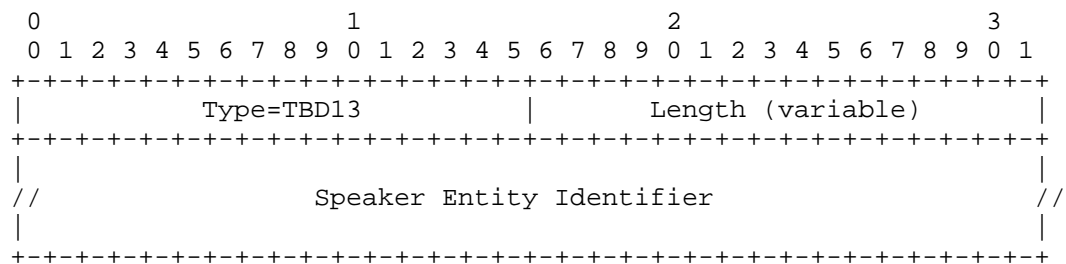


Figure 4: LSP-DB-VERSION TLV format

The type of the TLV is TBD5 and it has a fixed length of 8 octets. The value contains a 64-bit unsigned integer, carried in network byte order, representing the LSP State DB Version Number.

The 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 lifetime of the PCEP speaker. It identifies the PCEP speaker to its peers even if the speaker's IP address is changed.

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



The type of the TLV is TBD13 and it has a variable length, which MUST be greater than 0. The Value is padded to 4-octet alignment. The padding is not included in the Length field. 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 to make sure it remains unique.

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 this Speaker Entity Identifier.

If a PCEP speaker receives the SPEAKER-ENTITY-ID on a new PCEP session, that matches with an existing alive PCEP session, the PCEP speaker MUST send a PCERR with Error-type 20 Error-value TBD7 (suggested value - 7) 'Received an invalid Speaker Entity Identifier', and close the PCEP session.

4. Incremental State Synchronization

[I-D.ietf-pce-stateful-pce] describes the LSP state synchronization mechanism between PCCs and stateful PCEs. During the state synchronization, a PCC sends the information of all its LSPs (i.e., the full LSP-DB) to the stateful PCE. In order to reduce the state synchronization overhead when there is a small number of LSP state change in the network between PCEP session restart, this section defines a mechanism for incremental (Delta) LSP Database (LSP-DB) synchronization.

4.1. Motivation

According to [I-D.ietf-pce-stateful-pce], if a PCE restarts and its LSP-DB survived, PCCs with mismatched LSP State Database Version Number 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.

Figure 6 shows an example of LSP state synchronization.

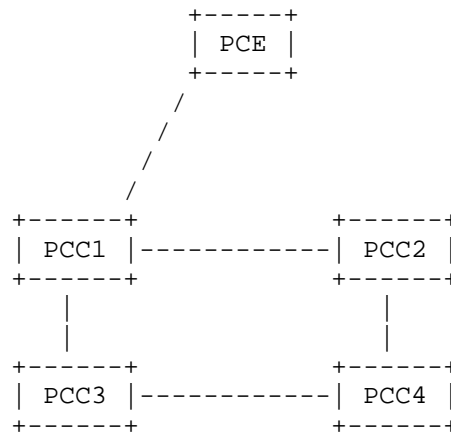


Figure 6: Topology Example

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 (e.g., in-band control channel for optical transport networks) 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/modified/deleted LSPs.

4.2. Incremental Synchronization Procedure

[I-D.ietf-pce-stateful-pce] describes state synchronization and Section 3 of this document, 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.

If both PCEP speakers include the LSP-DB-VERSION TLV in the OPEN object and the LSP-DB-VERSION TLV values match, the PCC MAY skip state synchronization. Otherwise, the PCC MUST perform state synchronization. Incremental State synchronization capability is advertised on a PCEP session during session startup using the DELTA-LSP-SYNC-CAPABILITY (D) bit in the capabilities TLV (see Section 7). Instead of dumping full LSP-DB to the stateful PCE again, the PCC synchronizes the delta (changes) as described in Figure 7 when D flag and S flag is set to 1 by both PCC and PCE. Other combinations of D and S flags setting by PCC and PCE result in full LSP-DB synchronization procedure as described in [I-D.ietf-pce-stateful-pce]. By setting the D flag to zero in the OPEN message, a PCEP speaker can skip the incremental synchronization optimization, resulting in a full LSP DB synchronization.

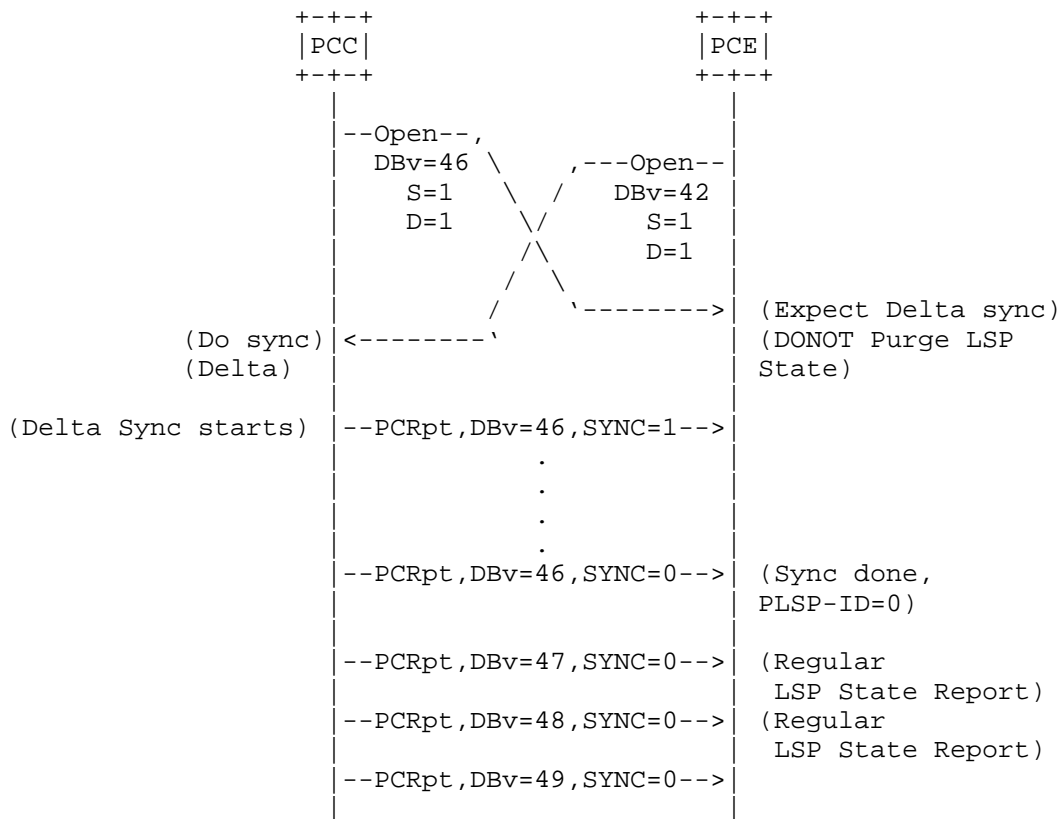


Figure 7: Incremental Synchronization Procedure

As per Section 3, the LSP State Database Version Number 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. The incremental state sync is done from the last LSP DB version received by the PCE to the latest DB version at the PCC. Note that the LSP State Database Version Number can wrap around, and in which case the incremental state sync would also wrap till the latest DB version number at the PCC.

In order to carry out incremental state synchronization, it is not necessary for a PCC to store a complete history of LSP Database

change for all time, but remember the LSP state changes (including LSP modification, setup and deletion), that the PCE did not get to process during the session down. Note that, a PCC would be unaware that a particular LSP report has been processed by the PCE before the session to PCE went down. So a PCC implementation MAY choose to store the LSP State Database Version Number with each LSP at the time its status changed, so that when a session is re-established an incremental synchronization can be attempted based on the PCE's last LSP State Database Version Number. For an LSP that is deleted at the PCC, the PCC implementation would need to remember the deleted LSP in some way to make sure this could be reported as part of incremental synchronization later. The PCC would discard this information based on a local policy, or when it determines that this information is no longer needed with sufficient confidence. In the example shown in Figure 7, the PCC needs to store the LSP state changes that happened between DB Version 43 to 46 and synchronizes these changes, when performing incremental LSP state update.

If a PCC finds out it does not have sufficient information to complete incremental synchronization after advertising incremental LSP state synchronization capability, it MUST send a PCErr with Error-Type 20 and Error-Value 5 'A PCC indicates to a PCE that it can not complete the state synchronization' (defined in [I-D.ietf-pce-stateful-pce]) and terminate the session. The PCC SHOULD re-establish the session with the D bit set to 0 in the OPEN message.

The other procedures and error checks remain unchanged from the full state synchronization ([I-D.ietf-pce-stateful-pce]).

5. PCE-triggered Initial Synchronization

5.1. Motivation

In networks such as optical transport networks, the control channel between network nodes can be realized through in-band overhead thus has limited bandwidth. With a stateful PCE connected to the network via one network node, it is desirable to control the timing of PCC state synchronization so as not to overload the low communication channel available in the network during the initial synchronization (be it incremental or full) when the session restarts, when there is comparatively large amount of control information needing to be synchronized between the stateful PCE and the network. The method proposed, i.e., allowing PCE to trigger the state synchronization, is similar to the function proposed in Section 6 but is used in different scenarios and for different purposes.

5.2. PCE-triggered Initial State Synchronization Procedure

Support of PCE-triggered initial state synchronization is advertised during session startup using the TRIGGERED-INITIAL-SYNC (F) bit in the STATEFUL-PCE-CAPABILITY TLV (see Section 7).

In order to allow a stateful PCE to control the LSP-DB synchronization after establishing a PCEP session, both PCEP speakers MUST set F bit to 1 in the OPEN message. If the LSP-DB-VERSION TLV is included by both PCEP speakers and the TLV value matches, the state synchronization can be skipped as described in Section 3.2. If the TLV is not included or the LSP-DB Version is mis-matched, the PCE can trigger the state synchronization process by sending a PCUpd message with PLSP-ID = 0 and SYNC = 1. The PCUpd message SHOULD include an empty ERO (with no ERO sub-object and object length of 4) as its intended path and SHOULD NOT include the optional objects for its attributes for any parameter update. The PCC MUST ignore such an update when the SYNC flag is set. If the TRIGGERED-INITIAL-SYNC capability is not advertised by a PCE and the PCC receives a PCUpd with the SYNC flag set to 1, the PCC MUST send a PCErr with the SRP-ID-number of the PCUpd, Error-Type 20 and Error-Value TBD4 (suggested value - 4) 'Attempt to trigger synchronization when the TRIGGERED-SYNC capability has not been advertised' (see Section 8.1). If the TRIGGERED-INITIAL-SYNC capability is advertised by a PCE and the PCC, the PCC MUST NOT trigger state synchronization on its own. If the PCE receives a PCRpt message before the PCE has triggered the state synchronization, the PCE MUST send a PCErr with Error-Type 20 and Error-Value TBD3 (suggested value - 3) 'Attempt to trigger synchronization before PCE trigger' (see Section 8.1).

In this way, the PCE can control the sequence of LSP synchronization among all the PCCs that are re-establishing PCEP sessions with it. When the capability of PCE control is enabled, only after a PCC receives this message, it will start sending information to the PCE. This PCE-triggering capability can be applied to both full and incremental state synchronization. If applied to the latter, the PCCs only send information that PCE does not possess, which is inferred from the LSP-DB version information exchanged in the OPEN message (see Section 4.2 for detailed procedure).

Once the initial state synchronization is triggered by the PCE, the procedures and error checks remain unchanged ([I-D.ietf-pce-stateful-pce]).

If a PCC implementation that does not implement this extension should not receive a PCUpd message to trigger state synchronization as per the capability advertisement, but if it were to receive it, it will behave as per [I-D.ietf-pce-stateful-pce].

6. PCE-triggered Re-synchronization

6.1. 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 re-synchronize this state even after the session has been 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.

6.2. PCE-triggered State Re-synchronization Procedure

Support of PCE-triggered state re-synchronization is advertised by both PCEP speakers during session startup using the TRIGGERED-RESYNC (T) bit in the STATEFUL-PCE-CAPABILITY TLV (see Section 7). The PCE can choose to re-synchronize its entire LSP database or a single LSP.

To trigger re-synchronization for an LSP, the PCE sends a Path Computation State Update (PCUpd) for the LSP, with the SYNC flag in the LSP object set to 1. The PCE SHOULD NOT include any parameter updates for the LSP, and the PCC MUST ignore such an update when the SYNC flag is set. The PCC MUST respond with a PCRpt message with the LSP state, SYNC Flag set to 0 and MUST include the SRP-ID-number of the PCUpd message that triggered the resynchronization. If the PCC cannot find the LSP in its database, PCC MUST also set the R (remove) flag [I-D.ietf-pce-stateful-pce] in the LSP object in the PCRpt message.

The PCE can also trigger re-synchronization of the entire LSP database. The PCE MUST first mark all LSPs in the LSP database that were previously reported by the PCC as stale and then send a PCUpd with an LSP object containing a PLSP-ID of 0 and with the SYNC flag set to 1. The PCUpd message MUST include an empty ERO (with no ERO sub-object and object length of 4) as its intended path and SHOULD NOT include the optional objects for its attributes for any parameter update. The PCC MUST ignore such update if the SYNC flag is set. 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 re-synchronization SHOULD be included in each of the PCRpt messages. If the PCC cannot re-synchronize the entire LSP database, the PCC MUST respond with PCErr message with Error-type 20 Error-value 5 'cannot complete the state synchronization' [I-D.ietf-pce-stateful-pce], and MAY terminate the session. The PCE MUST remove the stale mark for

the LSP that were previously reported by the PCC. Based on the local policy, the PCE MAY reattempt synchronization at a later time.

If the TRIGGERED-RESYNC capability is not advertised by a PCE 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 TBD4 (suggested value - 4) 'Attempt to trigger synchronization when the TRIGGERED-SYNC capability has not been advertised' (see Section 8.1).

Once the state re-synchronization is triggered by the PCE, the procedures and error checks remain unchanged from the full state synchronization ([I-D.ietf-pce-stateful-pce]). This would also include PCE triggering multiple state re-synchronization requests while synchronization is in progress.

If a PCC implementation that does not implement this extension should not receive a PCUpd message to trigger re-synchronization as per the capability advertisement, but if it were to receive it, it will behave as per [I-D.ietf-pce-stateful-pce].

7. Advertising Support of Synchronization Optimizations

Support for each of the optimizations described in this document requires advertising the corresponding capabilities during session establishment time.

The STATEFUL-PCE-CAPABILITY TLV is defined in [I-D.ietf-pce-stateful-pce]. This document defines following new flags in the STATEFUL-PCE-CAPABILITY TLV:

Bit	Description
TBD9 (suggested value 30)	S bit (INCLUDE-DB-VERSION)
TBD10 (suggested value 27)	D bit (DELTA-LSP-SYNC-CAPABILITY)
TBD11 (suggested value 26)	F bit (TRIGGERED-INITIAL-SYNC)
TBD12 (suggested value 28)	T bit (TRIGGERED-RESYNC)

If the S (INCLUDE-DB-VERSION) bit is set to 1 by both PCEP Speakers, the PCC will include the LSP-DB-VERSION TLV in each LSP Object. See Section 3.2 for details.

If the D (DELTA-LSP-SYNC-CAPABILITY) bit is set to 1 by a PCEP speaker, it indicates that the PCEP speaker allows incremental (delta) state synchronization. See Section 4.2 for details.

If the F (TRIGGERED-INITIAL-SYNC) bit is set to 1 by both PCEP Speakers, the PCE SHOULD trigger initial (first) state synchronization. See Section 5.2 for details.

If the T (TRIGGERED-RESYNC) bit is set to 1 by both PCEP Speakers, the PCE can trigger re-synchronization of LSPs at any point in the life of the session. See Section 6.2 for details.

See Section 8.3 for IANA allocations.

8. IANA Considerations

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

8.1. PCEP-Error Object

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

Error-Type	Meaning	Reference
6	Mandatory Object missing Error-Value= TBD1(suggested value 12): LSP-DB-VERSION TLV missing	[RFC5440] This document
20	LSP State synchronization error Error-Value= TBD2(suggested value 2): LSP Database version mismatch. Error-Value=TBD3(suggested value 3): Attempt to trigger synchronization before PCE trigger. Error-Value=TBD4(suggested value 4): Attempt to trigger a synchronization when the PCE triggered synchronization capability has not been advertised. Error-Value=TBD6(suggested value 6): Received an invalid LSP DB Version Number. Error-Value=TBD7(suggested value 7): Received an invalid Speaker Entity Identifier.	[I-D.ietf-pce-stateful-pce] This document This document This document This document This document This document

8.2. PCEP TLV Type Indicators

IANA is requested to make the following allocation in the "PCEP TLV Type Indicators" registry.

Value	Meaning	Reference
TBD5(suggested value 23)	LSP-DB-VERSION	This document
TBD13(suggested value 24)	SPEAKER-ENTITY-ID	This document

8.3. STATEFUL-PCE-CAPABILITY TLV

The STATEFUL-PCE-CAPABILITY TLV is defined in [I-D.ietf-pce-stateful-pce] and a registry is requested to be created to manage the flags in the TLV. IANA is requested to make the following allocation in the aforementioned registry.

Bit	Description	Reference
TBD11 (suggested value 26)	TRIGGERED-INITIAL-SYNC	This document
TBD10 (suggested value 27)	DELTA-LSP-SYNC-CAPABILITY	This document
TBD12 (suggested value 28)	TRIGGERED-RESYNC	This document
TBD9 (suggested value 30)	INCLUDE-DB-VERSION	This document

9. Manageability Considerations

All manageability requirements and considerations listed in [RFC5440] and [I-D.ietf-pce-stateful-pce] apply to PCEP protocol extensions defined in this document. In addition, requirements and considerations listed in this section apply.

9.1. Control of Function and Policy

A PCE or PCC implementation MUST allow configuring the state synchronization optimization capabilities as described in this document. The implementation SHOULD also allow the operator to configure the Speaker Entity Identifier (Section 3.3.2). Further, the operator SHOULD be to be allowed to trigger the re-synchronization procedures as per Section 6.2.

9.2. Information and Data Models

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

9.3. Liveness Detection and Monitoring

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

9.4. Verify Correct Operations

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

9.5. Requirements On Other Protocols

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

9.6. Impact On Network Operations

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

The state synchronization optimizations described in this document can result in a reduction of the amount of data exchanged and the time taken for a stateful PCE to be fully operational when a PCEP session is re-established. The ability to trigger re-synchronization by the PCE can be utilized by the operator to sanity check its state and recover from any mismatch in state without tearing down the session.

10. Security Considerations

The security considerations listed in [I-D.ietf-pce-stateful-pce] apply to this document as well. However, this document also introduces some new attack vectors. An attacker could spoof the SPEAKER-ENTITY-ID and pretend to be another PCEP speaker. An attacker may flood the PCC with triggered re-synchronization request at a rate which exceeds the PCC's ability to process them, either by spoofing messages or by compromising the PCE itself. The PCC can respond with PCErr message as described in Section 6.2 and terminate the session. Thus securing the PCEP session using Transport Layer Security (TLS) [I-D.ietf-pce-pceps], as per the recommendations and best current practices in [RFC7525], is RECOMMENDED. An administrator could also expose the speaker entity id as part of the certificate, for the peer identity verification.

11. Acknowledgments

We would like to thank Young Lee, Sergio Belotti and Cyril Margaria for their comments and discussions.

Thanks to Jonathan Hardwick for being the document shepherd and provide comments and guidance.

Thanks to Tomonori Takeda for Routing Area Directorate review.

Thanks to Adrian Farrel for TSVART review and providing detailed comments and suggestions.

Thanks to Daniel Franke for SECDIR review.

Thanks to Alvaro Retana, Kathleen Moriarty, and Stephen Farrell for comments during the IESG evaluation.

Thanks to Deborah Brungard for being the responsible AD and guiding the authors as needed.

12. 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

13. References

13.1. Normative References

- [I-D.ietf-pce-stateful-pce]
Crabbe, E., Minei, I., Medved, J., and R. Varga, "PCEP Extensions for Stateful PCE", draft-ietf-pce-stateful-pce-18 (work in progress), December 2016.
- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <<http://www.rfc-editor.org/info/rfc2119>>.
- [RFC5440] Vasseur, JP., Ed. and JL. Le Roux, Ed., "Path Computation Element (PCE) Communication Protocol (PCEP)", RFC 5440, DOI 10.17487/RFC5440, March 2009, <<http://www.rfc-editor.org/info/rfc5440>>.

13.2. Informative References

- [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>>.
- [RFC8051] Zhang, X., Ed. and I. Minei, Ed., "Applicability of a Stateful Path Computation Element (PCE)", RFC 8051, DOI 10.17487/RFC8051, January 2017, <<http://www.rfc-editor.org/info/rfc8051>>.
- [I-D.ietf-pce-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, W., and D. Dhody, "Secure Transport for PCEP", draft-ietf-pce-pceps-11 (work in progress), January 2017.

Authors' Addresses

Edward Crabbe
Oracle

E-Mail: edward.crabbe@gmail.com

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

E-Mail: 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

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
Divyashree Techno Park, Whitefield
Bangalore, Karnataka 560066
India

EMail: dhruv.ietf@gmail.com

PCE Working Group
Internet Draft
Intended Status: Standard Track

Y. Lee
Huawei
G. Bernstein
Grotto Networking
H. Zheng
D. Dhody
Huawei

Expires: January 2015

July 2, 2014

PCEP Extensions in Support of Transporting Traffic Engineering Data

draft-lee-pce-transporting-te-data-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 January 2, 2009.

Copyright Notice

Copyright (c) 2014 IETF Trust and the persons identified as the document authors. All rights reserved.

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

Abstract

In order to compute and provide optimal paths, Path Computation Elements (PCEs) require an accurate and timely Traffic Engineering Database (TED). Traditionally this TED has been obtained from a link state routing protocol supporting traffic engineering extensions. This document discusses possible alternatives to TED creation. This document gives architectural alternatives for these enhancements and their potential impacts on network nodes, routing protocols, and PCE.

Table of Contents

1. Introduction.....	3
1.1. TED Creation and Maintenance via IGP-TEs.....	4
2. Alternative TED Creation & Maintenance for a PCE.....	6
2.1. Architecture Options.....	7
2.1.1. Nodes Send TE Info to all PCEs.....	12
2.1.2. Nodes Send TE Info via an Intermediate System.....	12
2.1.3. Nodes Send TE Info to At Least One PCE.....	12
2.2. Nodes Finding PCEs.....	13
2.3. Node TE Information Update Procedures.....	14
2.4. PCE TED Maintenance Procedures.....	14
3. Standardization and Protocol Considerations.....	14
3.1. Architecture Specific Standardization Aspects.....	15
4. Security Considerations.....	16
5. IANA Considerations.....	16
6. Conclusions.....	17
7. Acknowledgments.....	17
8. References.....	17
8.1. Normative References.....	17
8.2. Informative References.....	18
Author's Addresses.....	19
Intellectual Property Statement.....	19
Disclaimer of Validity.....	20

1. Introduction

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

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

Section 4.3 of [RFC4655] describes the potential load of the TED on a network node and proposes an architecture where the TED is maintained by the PCE rather than the network nodes. However, it does not describe how a PCE would obtain the information needed to populate its TED. PCE may construct its TED by participating in the IGP ([RFC3630] and [RFC5305] for MPLS-TE; [RFC4203] and [RFC5307] for GMPLS). An alternative is offered by BGP-LS [BGP-LS].

In this document we propose approaches for creating and maintaining the TED directly on a PCE as an alternative to IGPs and BGP flooding and investigate the impact from the PCE, routing protocol, and node perspectives.

New application areas for GMPLS and PCE in optical transport networks include Wavelength Switched Optical Networking (WSON) and Optical Transport Networks (OTN). WSON scenarios can be divided into routing wavelength assignment (RWA) problems where PCE requires detailed information about switching node asymmetries and wavelength constraints as well as detailed up to date information on wavelength usage per link [WSON-Frame]. As more data is anticipated to be made available to PCE with addition of OTN [RFC7062] and Flexi-grid [Flexi-grid] and possible with some optical impairment data [WSON-IMP-Info] even with the minimum set specified in [G.680], the total amount of data requires significantly more information to be held in the TED than is required for other traffic engineered networks.

In some circumstances such additional information could "bog down" the routing protocols on the nodes from a data processing, a storage, or communications perspective. In environments where PCEs

are external to the nodes running the routing protocol, and where the information in the TED is not used by the switching nodes it makes sense to investigate alternative methods to create and maintain the TED at its place of use, i.e., the PCE.

Recent development of a stateful PCE Model [PCE-Initiated] changes the PCE operation from path computation alone to include the support of PCE-initiated LSPs. With a stateful PCE model, it is also noted that LSP-DB is maintained by the PCE. 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) [S-PCE-GMPLS]. To support all these recent changes in a stateful PCE model, a direct PCE interface to each PCC has to be supported. Relevant TED information can also be transported from each node to PCE using this PCC-PCE interface. Any resource changes in the node and links can also be quickly updated to PCE using this interface. Convergence time of IGP in GMPLS networks may not be quick enough to support on-line dynamic connectivity required for some applications.

This draft does not advocate that the alternative methods specified in this draft should completely replace the IGP-TE or BGP-LS as the method of creating the TED. The split between the data to be distributed via an IGP and the information conveyed via one of the alternatives in this document depends on the nature of the network situation. One could potentially choose to have some traffic engineering information distributed via an IGP while other more specialized traffic information is only conveyed to the PCEs via an alternative interface discussed here. In addition, the methods specified in this draft is only relevant to a set of architecture options where routing decisions are wholly or partially made in the PCE.

However, the networks that do not support IGP-TE/BGP-LS, the method proposed by this draft may be very relevant.

1.1. TED Creation and Maintenance via IGP-TEs

Routing protocols, in particular, IGP-TEs such as Open Shortest Path First (OSPF) and Intermediate system to intermediate system (IS-IS), take on a number of roles with respect to the control and data planes for IP, MPLS, and GMPLS. In all three technology families the underlying control plane communications technology is IP and hence all utilize the IGPs ability to control and run the IP data plane.

For the IP layer, the IGP directly establishes data plane connectivity. In the MPLS and GMPLS cases separate signaling protocols are used to directly control the data plane connectivity and in these cases the prime purpose of the routing protocol is to furnish network topology and resource status information used by path computation algorithms on the nodes or PCEs. Hence in the IP case the IGP is directly service impacting, while in the MPLS/GMPLS case it is only indirectly service impacting.

The IP layer information and the MPLS/GMPLS data plane layer information may be kept by the IGPs in two different information stores. These are referred to as databases but are not necessarily relational databases. In OSPF the information directly related to IP connectivity (and hence the control communications plane for all three technologies) and non-IP advertisements are kept in the link state database (LSDB), while information related to traffic engineering used by MPLS and GMPLS is kept in a (conceptually) separate TED which can be considered a subset of the LSDB. This TED information is distributed in a different data structure (Opaque LSA [RFC5250]). When we talk about adding additional technology-specific GMPLS information used for path computation we are only talking about adding to the TED and not the IP portion of the LSDB.

There are three main functions performed by an IGP: (a) hello protocol, (b) database synchronization (with neighbors), (c) database updates.

Data Plane Technologies	Hello Protocol	Database Sync & Updates
IP	Establish Control & Data Plane Adjacencies	LSDB
MPLS	Establish Control & Data Plane Adjacencies	LSDB & TED
GMPLS	Establish Control Plane Adjacencies (only)	LSDB & TED

Table 1 Main Functions of an IGP for various technologies

The procedures for maintaining LSDBs and TEDs in IGP-TEs have been very successful and well proven over time. These consist of:

1. Ageing the individual pieces of information in the TED (including discarding them when the information gets too old) to remove stale information from the TED.
2. Originator of the information being required to periodically resend TED information to prevent it from being discarded.
3. Originator of the information sending updates of information as needed, but subject to limits on how many/often these can be sent to keep the TED up-to-date, but to avoid swamping the network.
4. Reliable method for getting this information to other peers (flooding) to ensure that the information is delivered to all participants.
5. An efficient database synchronization mechanism for sharing info with a newly established peer.

From a PCE perspective, however, participating in an IGP, even as a passive receiver of IGP information, can place a significant load on the PCE. The IGP can be quite "chatty" when there are frequent updates to the use of the network, meaning that the PCE must dedicate significant processing to parsing protocol messages and updating the TED. Furthermore, to be truly useful, a PCE implementation would need to support OSPF and IS-IS.

2. Alternative TED Creation & Maintenance for a PCE

Given that nodes, by their position and role in the network, have accurate traffic engineering information concerning their local link ends and switching properties, it seems natural that, if other nodes in the network cannot make use of this information or do not want it, the information should only be conveyed to interested PCEs. In such case the flooding of TE information to all nodes may not be very efficient in terms of memory, CPU, bandwidth, etc.

The benefits of such an approach include:

- o Node: reduced storage demands (doesn't keep the entire TED)
- o Node: reduced processing demands for TED updates and synchronization

- o Control Plane: reduced overall communication demands since the TED is not being updated and maintained on all nodes in the network.
- o PCE: More timely TED updates are possible.
- o Information distribution constraints, such as seen in [Imp-Frame] can be met.

To quantify the previous advantages requires a bit more detail on how such an approach could actually be accomplished. The key pieces needed to implement such an approach include:

- o Multiple PCEs must be supported for robustness and load sharing.
- o Nodes must be able to find a PCE to which to send their traffic engineering information.
- o Nodes must have procedures and a mechanism (protocols) with which to communicate their TE information to a PCE. PCEs must have procedures and a mechanism (protocols) with which to receive this TE information from nodes.
- o Efficient mechanisms must exist in the multi-PCE case to ensure all PCEs have the same TED.

The advantages of using an alternative to IGP-TE comes at the cost of:

- o Additional protocols to be configured and secured. Recall that we still must have an IP IGP for control plane communications.
- o Any new protocols/implementations for alternative TED creation still must support many IGP-TE like features such as removal of stale information, reliable delivery of updates to all participants, recovery after reboots/crashes/upgrades, etc. It should also work along with IGP-TE/BGP-LS TED mechanism with some information in the TED received from existing mechanisms.
- o Node mechanisms to discover PCEs that are capable and willing to accept direct TED updates.

2.1. Architecture Options

There are three general architectural alternatives based on how nodes get their local TED information to the PCEs: (1) Nodes send local information to all PCEs; (2) Nodes send local information to

an intermediate server that will send to all PCEs; (3) Nodes send local information to at least one PCE and have the PCEs share this information with each other. An important functionality that needs to be addressed in each of these approaches is how a new PCE gets initialized in a reasonably timely fashion.

Figures 1-3 show examples of three options for nodes to share local TED information with multiple PCEs. As in the IGP case we assume that switching nodes know their local properties and state including the state of all their local links. In these figures the data plane links are shown with the character "o"; TE information flow from nodes to PCE by the characters "|", "-", "/", or "\"; and PCE to PCE TE information, if any, by the character "i".

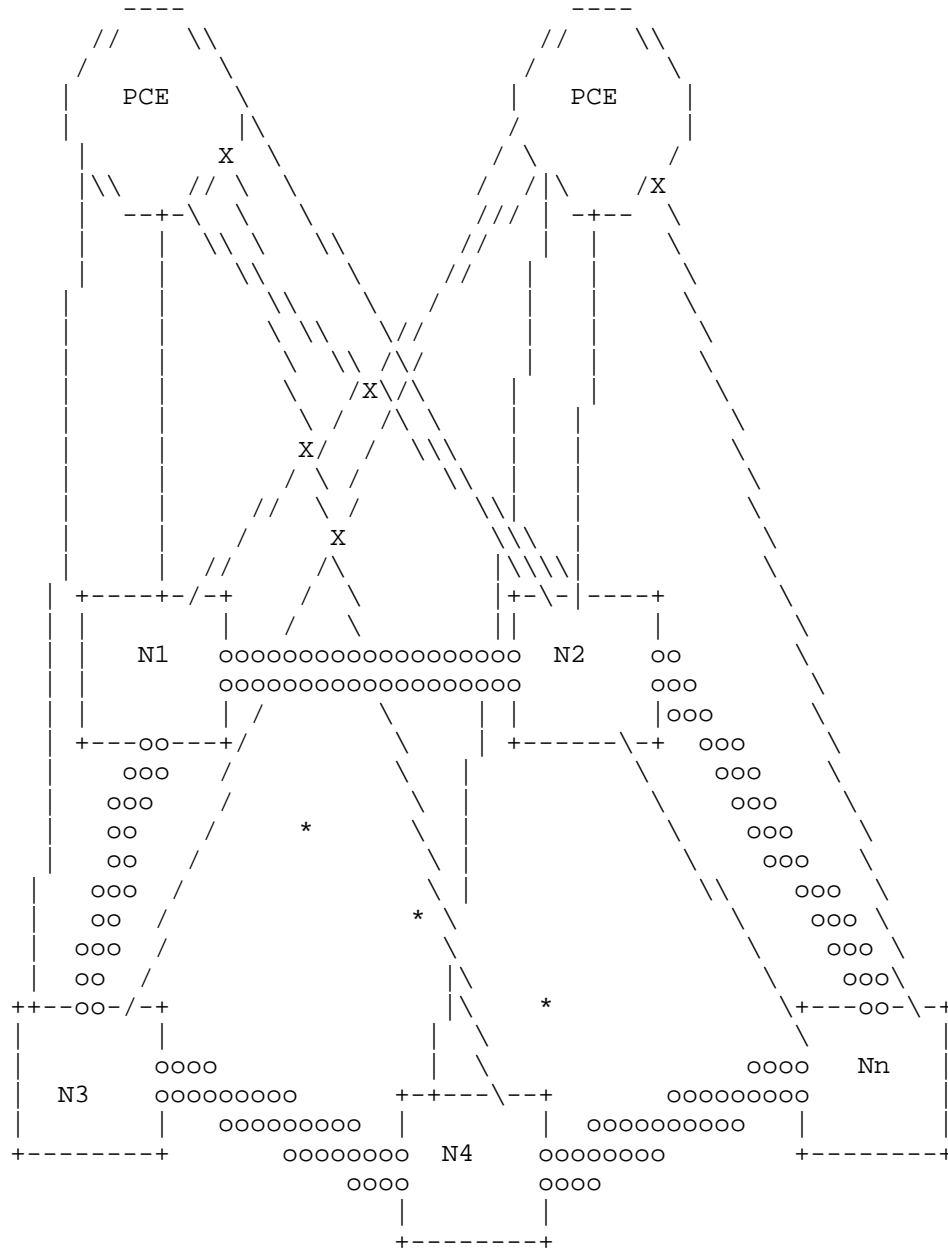


Figure 1 . Nodes send local TE information directly to all PCEs

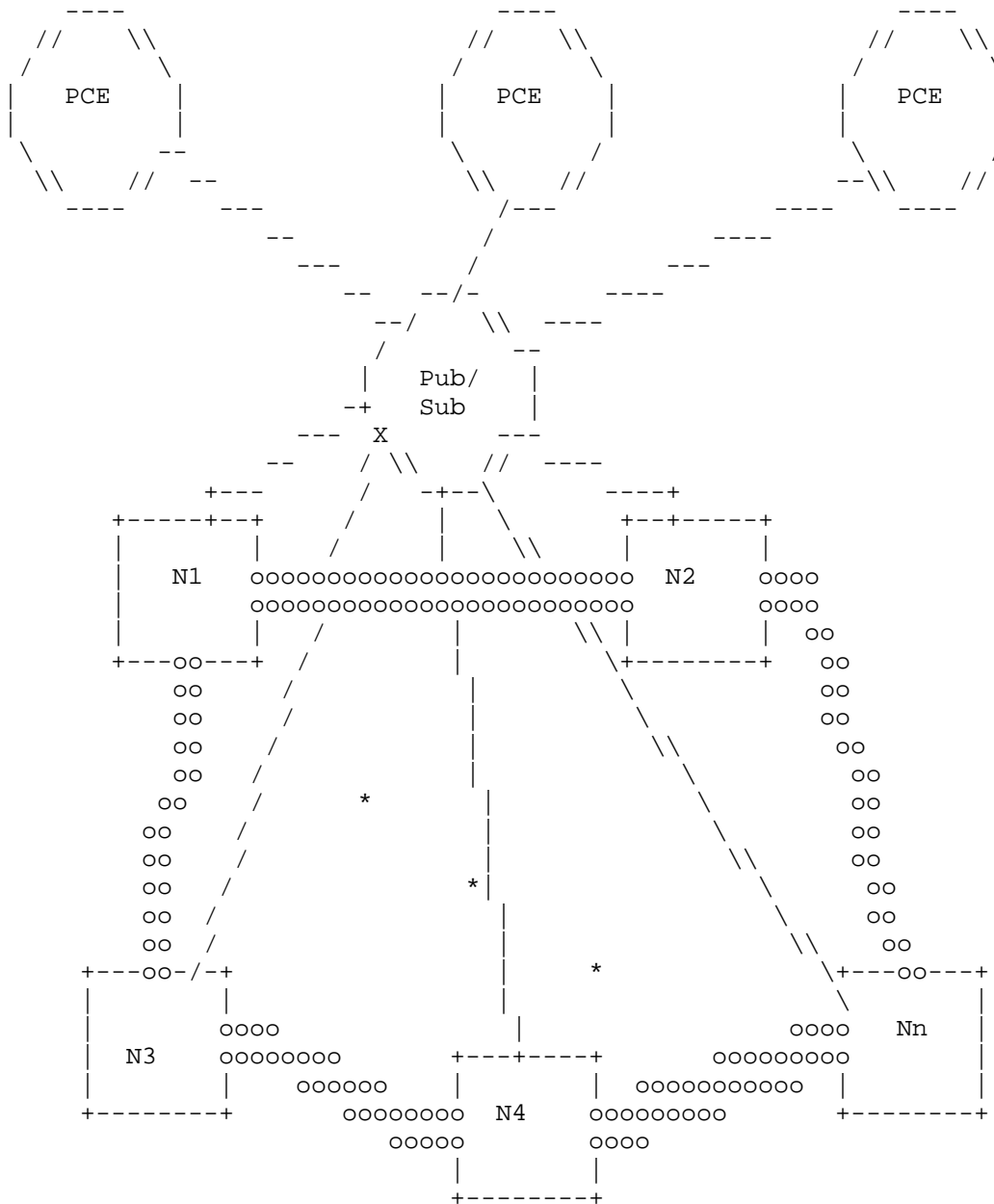


Figure 2 . Nodes send local TE information to PCEs via an intermediary (publish/subscribe)server

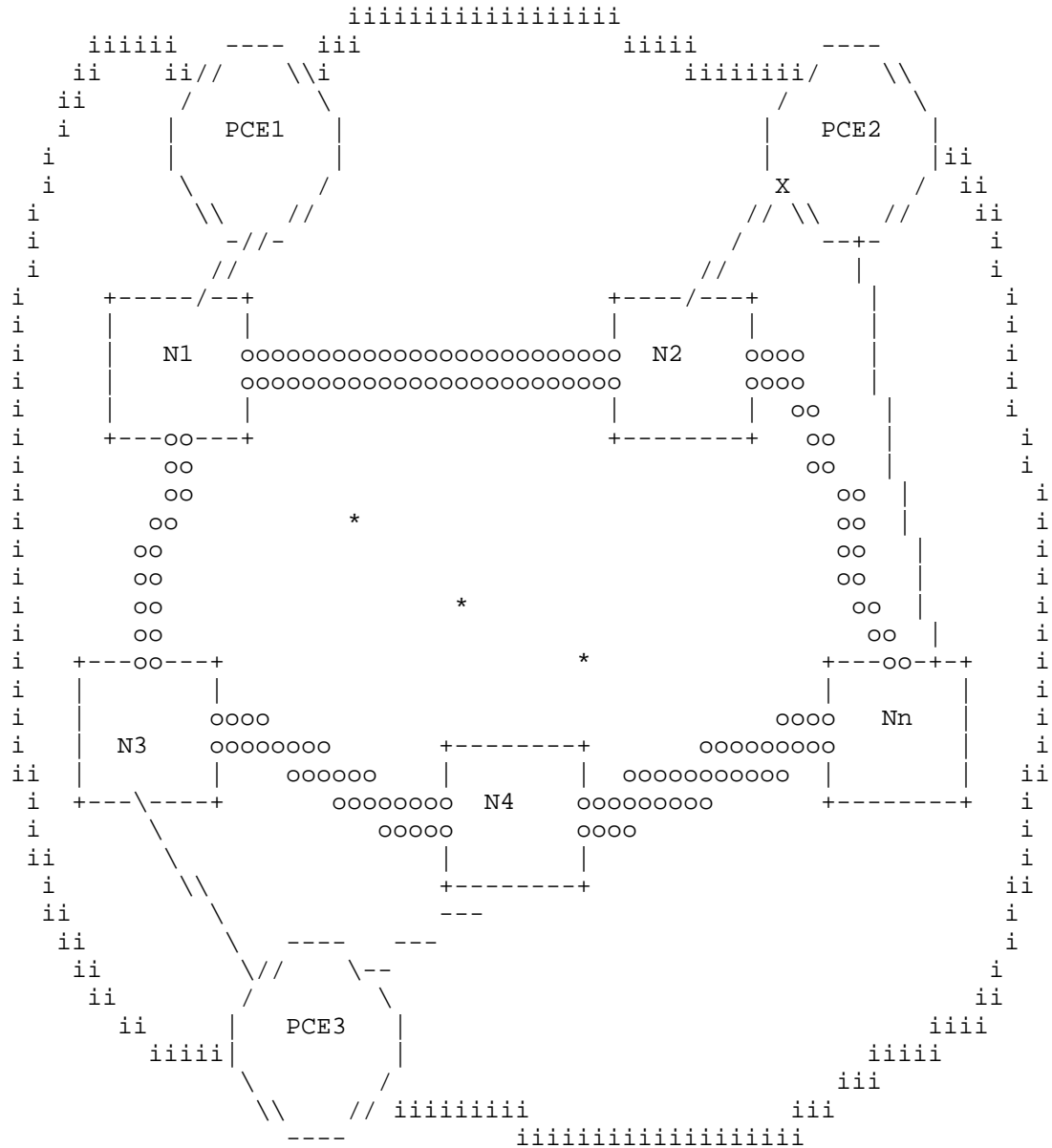


Figure 3 . Nodes send local TE information to at least one PCE and have the PCEs share TED information

2.1.1. Nodes Send TE Info to all PCEs

Architectural alternative 1 shown in Figure 1, illustrates nodes sending their local TE information to all PCEs within there domain. As the number of PCEs grow we have scaling concerns. However, if we are only talking about 2-3 PCEs, then we do not have this scaling concern. In particular each node needs to keep track of which PCE it has sent information to and update that information.

If a new PCE is added to the domain the node must send all its local TED information to that PCE rather than just sending status updates.

2.1.2. Nodes Send TE Info via an Intermediate System

Architecture alternative 2 is shown in Figure 2. This architecture reduces the burden on switching nodes by having the nodes send TE information to an intermediate system. This general approach is typically described in the software literature as a publish/subscribe paradigm. Here the nodes send their local TED information to an intermediate entity whose job is to insure that all PCEs receive this information. The nodes in this case being the publishers of the information and the PCEs the subscribers of the information. Publish/subscribe functionality can be found in general messaging oriented middleware such as the Java Messaging Service [JMS] and many others. A routing specific example of this approach is seen in BGP route reflectors [RFC4456].

Note that the publish/subscribe entity can be collocated with a PCE. This would then looks like a master/slave type system architecture.

If a new PCE is added then the intermediate server will need to work with this new PCE to initialize its TED. Hence the publish/subscribe entity will need to also keep a copy of the entire TED and for reliability purposes a redundant server would be required. The publish/subscribe entity itself can be a PCE.

Architecture alternative 2 could be useful when there are a number of PCEs in the network and as such there is the scaling issue with each of the NEs talking to all the PCEs. The advantage of this alternative would diminish when we are dealing only with only a few PCEs.

2.1.3. Nodes Send TE Info to At Least One PCE

In this architectural alternative, shown in Figure 3, each node would be associated with at least one PCE. This implies that each

PCE will only have partial TED information directly from the nodes. It would be the responsibility of a node to get its local TED information to its associated PCE, then the PCEs within a domain would then need to share the partial TED information they learned from their associated nodes with each other so that they can create and maintain the complete TED. As we have seen in section 1.1. this is very similar to part of the functionality provided by a link state protocol, but in this case the protocol would be used between PCEs so that they can share the information they have obtained from their associated switching nodes (rather than from attached links as in a regular link state protocol).

To allow for this sharing of information PCEs would need to peer with each other. PCE discovery extensions [RFC4674] could be used to allow PCEs to find other PCEs. If a new PCE is added to the domain it would need to peer with at least one other PCE and then link state protocol procedures for TED synchronization could then be used to initialize the new PCEs TED.

A number of approaches can be used to ensure control plane resilience in this architecture. (1) Each node can be configured with a primary and a secondary PCE to send its information to; In case of failure of communications with the primary PCE the node would send its information to a secondary PCE (warm standby). (2) Each node could be configured to send its information to two different PCEs (hot standby).

2.2. Nodes Finding PCEs

In cases 1 and 3 nodes need to send TE information directly to PCEs. Path Computation Clients (PCCs) and network nodes participating in an IGP (with or without TE extensions) have a mechanism to discover a PCE and its capabilities. [RFC4674] outlines the general requirements for this mechanism and extensions have been defined to provide information so that PCCs can obtain key details about available PCEs in OSPF [RFC5088] and in IS-IS [RFC5089].

After finding candidate PCEs, a node would need to see which if any of the PCEs actually want to receive TE information directly from this node.

In architectural alternative 2 (publish/subscribe) the location of intermediate system would either need to be configured or PCE discovery could be extended so that when a node asks a PCE if it wants to hear TE info the PCE points it to the intermediate publish/subscribe system.

2.3. Node TE Information Update Procedures

First a node must establish an association between itself and a PCE or intermediate system that will be maintaining a TED. It is the responsibility of the node to share TE information concerning its local environment, e.g., links and node properties. General and technology specific information models would specify the content of this information while the specific protocols would determine the format. Note that a node would not be sending to the PCE information it might be passed from neighbor nodes. Note that data plane neighbor information would be passed to the PCE embedded in TE link information.

There will be cases where the node would have to send to the PCE only a subset of TE link information depending on the path computation option. For instance, if the node is responsible for routing while the PCE is responsible for wavelength assignment for the route, the node would only need to send the PCE the WSON link usage information. This path computation option is referred to as separate routing (R) and wavelength assignment (WA) option in [PCE-WSON].

2.4. PCE TED Maintenance Procedures

The PCE is responsible for creating and maintaining the TED that it will use. Key functions include:

1. Establishing and authenticating communications between the PCE and sources of TED information.
2. Timely updates of the TED with information received from nodes, peers or other entities.
3. Verifying the validity of information in the TED, i.e., ensure that the network information obtained from nodes or elsewhere is relatively timely, or not stale.

3. Standardization and Protocol Considerations

In the previous section we examined a number of architectural alternatives for TED creation and maintenance on a PCE. Here we examine aspects of these alternatives that could be suitable for standardization. First there are a number of items and functions that can be independent of the particular architectural alternatives used, these include:

- o An information model for the TED

- o Basic PCE TED creation and maintenance procedures
- o Information packaging for use in TED creation, maintenance and exchange
- o NE to PCE (or Pub/Sub) communication of TED information --- interface and protocol (e.g. PCEP)
- o NEs discovering PCE (or Pub/Sub) for TED creation and maintenance purposes

By the "information model" for the TED we mean the raw information that a path computation algorithm would work with somewhat independent of how it might be packaged for TED maintenance and creation. Initial efforts along these lines have started at CCAMP for wavelength switched optical networks for non-impairment RWA [WSON-Info] and impairment aware RWA [WSON-IMP-Info].

Given a TED information model if we can agree on basic PCE TED creation and maintenance procedures we can then come up with a standardized way to package the information for use in such procedures. The analogy here is with an IGP's database maintenance procedures such as aging and the packaging of link state information into LSA (link state advertisements). LSAs form the basic chunks of an IGP's database. OSPF LSAs include an age field to assist in the ageing procedure and also has an advertising router field that aids in redistribution decisions, i.e., flooding. However the detailed TE information is encoded in LSAs via type length value (TLV) structures and it is this information that is used in path computation.

From there we could standardize the interface between a NE and a PCE for communication of TE information. This interface includes NE and PCE behaviors as well as a communications protocol.

Finally for the common behaviors we need a way for the NEs to find the PCEs or an intermediate publish/subscribe system to which they will send their TE information. As was previously pointed out this could be based on small enhancements to existing PCE discovery mechanisms.

3.1. Architecture Specific Standardization Aspects

Case 1: NEs send to all PCEs

This case has commonalities with both cases 2 and 3 and does not appear to have unique standardization aspects. As pointed out in section 2.1. we do need to consider when a new PCE comes online.

Case 2: Publish/Subscribe Server

In this case we would need to additionally standardize

1. how a new PCE coming online synchronizes with the publish/subscribe server
1. how PCEs and publish subscribe server communicate
2. Redundancy for publish subscribe server

Case 3: PCE to PCE sharing TE information learned from NEs

Here we would need the following additional mechanisms standardized:

1. The PCE to PCE interface and protocol
2. The method for PCEs to discover PCEs for the purpose of TE information sharing
3. PCE to PCE association for information sharing, in particular sharing update information.

4. Security Considerations

This draft discusses an alternative technique for PCEs to build and maintain a traffic engineering database. In this approach network nodes would directly send traffic engineering information to a PCE. It may be desirable to protect such information from disclosure to unauthorized parties in addition it may be desirable to protect such communications from interference (modification) since they can be critical to the operation of the network. In particular, this information is the same or similar to that which would be disseminated via a link state routing protocol with traffic engineering extensions.

5. IANA Considerations

This version of this document does not introduce any items for IANA to consider.

6. Conclusions

This document introduced several alternative architectures for PCEs to create and maintain a traffic engineering database (TED) via information directly or indirectly received from network elements and identified common aspects of these approaches. The TED is a critical piece of the overall PCE architecture since without it path computations cannot proceed. Though not explicitly out of scope the PCE working group does not have a work item or study item devoted to TED creation and maintenance. Such a work item can lead to enhanced interoperability and simplicity of PCE implementations. This document identified several common areas within these alternatives that could be standardized. In addition, the alternative approaches to TED creation and maintenance discussed here offloads both the network nodes and routing protocols from either some or all TED creation and maintenance duties at the same time it does not add significant new processing to a PCE that has already been participating in IGP based TED creation and maintenance.

7. Acknowledgments

TDB.

8. References

8.1. Normative References

- [RFC4655] Farrel, A., Vasseur, J.-P., and J. Ash, "A Path Computation Element (PCE)-Based Architecture", RFC 4655, August 2006.
- [RFC4674] Le Roux, J., Ed., "Requirements for Path Computation Element (PCE) Discovery", RFC 4674, October 2006.
- [RFC5088] Le Roux, JL., Ed., Vasseur, JP., Ed., Ikejiri, Y., and R. Zhang, "OSPF Protocol Extensions for Path Computation Element (PCE) Discovery", RFC 5088, January 2008.
- [RFC5089] Le Roux, JL., Ed., Vasseur, JP., Ed., Ikejiri, Y., and R. Zhang, "IS-IS Protocol Extensions for Path Computation Element (PCE) Discovery", RFC 5089, January 2008.
- [RFC5250] Berger, L., Bryskin, I., Zinin, A., and R. Coltun, "The OSPF Opaque LSA Option", RFC 5250, July 2008.

8.2. Informative References

- [JMS] Java Message Service, Version 1.1, April 2002, Sun Microsystems.
- [PCE-Initiated] E. Crabbe, et. al., "PCEP Extensions for PCE-initiated LSP Setup in a Stateful PCE Model", draft-ietf-pce-pce-initiated-lsp, work in progress.
- [S-PCE-GMPLS] X. Zhang, et. al, "Path Computation Element (PCE) Protocol Extensions for Stateful PCE Usage in GMPLS-controlled Networks", draft-ietf-pce-pcep-stateful-pce-gmpls, work in progress.
- [BGP-LS] H. Gredler, et. al., "North-Bound Distribution of Link-State and TE Information using BGP", draft-ietf-idr-ls-distribution, work in progress.
- [Flexi-grid] O. Gonzalez de Dios, Ed., and R. Casellas, Ed., "Framework and Requirements for GMPLS based control of Flexi-grid DWDM networks", draft-ietf-ccamp-flexi-grid-fwk, work-in-progress.
- [PCE-WSN] Y. Lee, G. Bernstein, "PCEP Requirements for the support of Wavelength Switched Optical Networks (WSN)", work in progress, draft-lee-pce-wson-routing-wavelength-05.txt, February 2009.
- [RFC4456] Bates, T., Chen, E., and R. Chandra, "BGP Route Reflection: An Alternative to Full Mesh Internal BGP (IBGP)", RFC 4456, April 2006.
- [Imp-Frame] G. Bernstein, Y. Lee, D. Li, A Framework for the Control and Measurement of Wavelength Switched Optical Networks (WSN) with Impairments, Work in Progress, October 2008.
- [WSN-Frame] Y. Lee, G. Bernstein, W. Imajuku, "Framework for GMPLS and PCE Control of Wavelength Switched Optical Networks", work in progress: draft-ietf-ccamp-wavelength-switched-framework-01.txt, February 2009.
- [WSN-IMP-Info] Y. Lee, G. Bernstein, "Information Model for Impaired Optical Path Validation", work in progress: draft-bernstein-wson-impairment-info-02.txt, March 2009.

Author's Addresses

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

Phone: (469) 277-5838
Email: leeyoung@huawei.com

Greg Bernstein
Grotto Networking

Email: gregb@grotto-networking.com

Haomian Zheng
Huawei Technologies Co., Ltd.
F3-1-B R&D Center, Huawei Base,
Bantian, Longgang District
Shenzhen 518129 P.R.China

Phone: +86-755-28979835
Email: zhenghaomian@huawei.com

Dhruv Dhody
Huawei Technologies
Leela Palace
Bangalore, Karnataka 560008
INDIA

Email: dhruv.ietf@gmail.com

Contributor's Addresses

Intellectual Property Statement

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.

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.

Acknowledgment

Funding for the RFC Editor function is currently provided by the Internet Society.

Network Working Group
Internet-Draft
Intended status: Informational
Expires: January 2, 2015

V. Lopez
O. Gonzalez de Dios
Telefonica I+D
D. King
Old Dog Consulting
S. Previdi
Cisco Systems, Inc.
J. Tantsura
Ericsson
July 1, 2014

Traffic Engineering Database dissemination for Hierarchical PCE
scenarios
draft-lopez-pce-hpce-ted-02

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 January 2, 2015.

Copyright Notice

Copyright (c) 2014 IETF Trust and the persons identified as the document authors. All rights reserved.

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

Table of Contents

1. Introduction	2
1.1. Parent PCE Domain Topology	3
1.2. Parent PCE TED requirements	4
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	9
4.1.1. Node Descriptors	9
4.1.2. Link Descriptors	9
4.1.3. Mapping OSPF TE parameters into BGP-LS attribute . .	10
4.2. Mapping from ISIS-TE	10
5. Manageability Considerations	10
6. Security Considerations	10
7. References	10
7.1. Normative References	10
7.2. Informative References	11
Authors' Addresses	11

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] 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]. For instance, [RFC6805] and [I-D.draft-ietf-pce-inter-area-as-applicability] 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. [I-D.draft-ietf-idr-te-pm-bgp] extends BGP-LS to disseminate traffic engineering information. The parameters considered are: delay, packet loss and bandwidth.

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]. 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] 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.

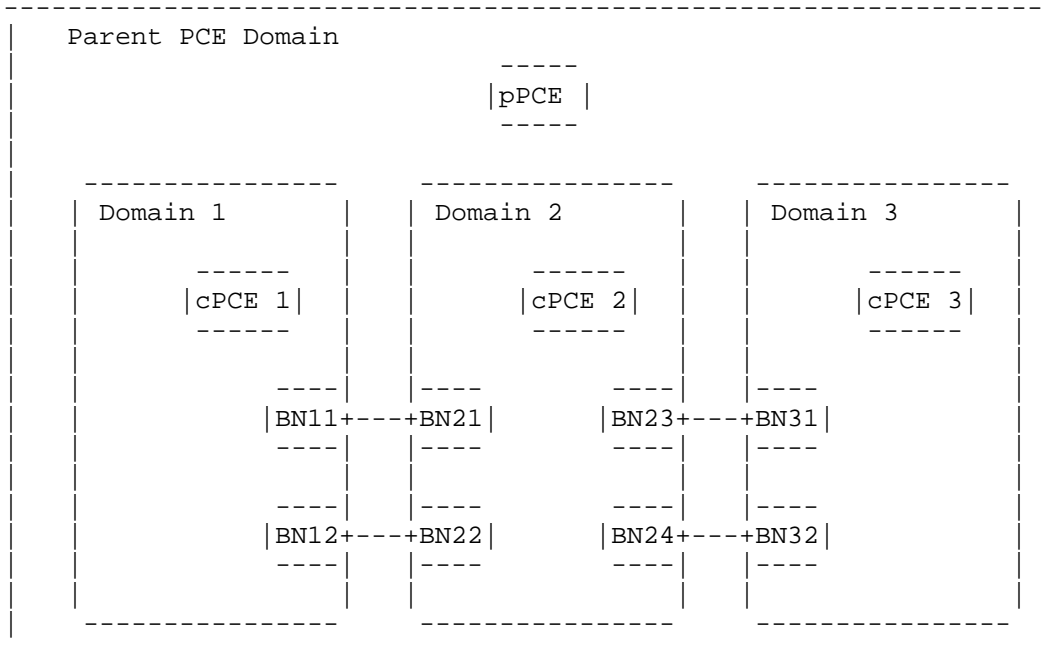


Figure 1: Example of Hierarchical PCE architecture

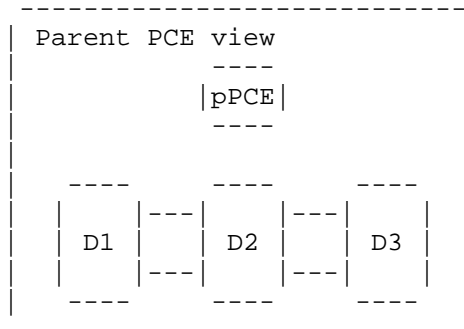


Figure 2: Parent PCE topology information

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 some intra-domain information for virtual or candidate paths with some TE information. 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.

BGP-LS [I-D.draft-ietf-idr-ls-distribution] 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]). 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.

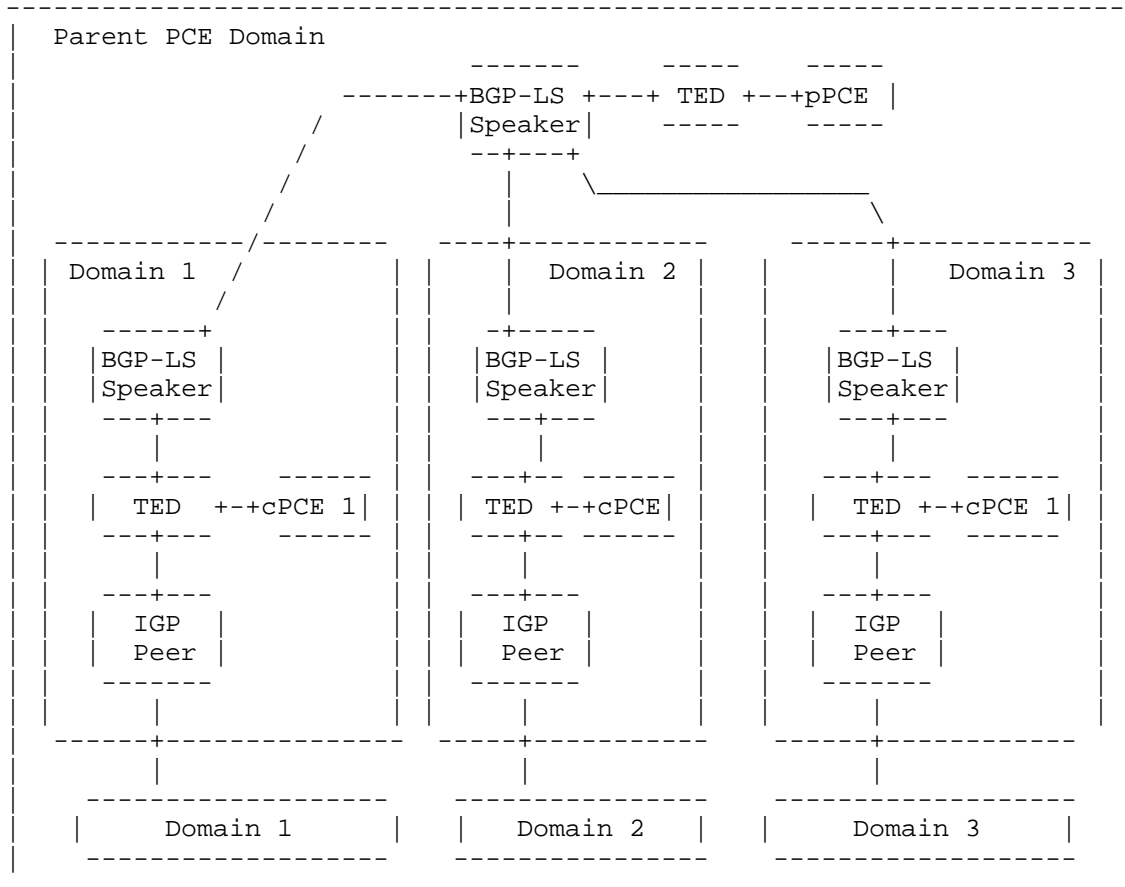


Figure 3: Example of Hierarchical PCE architecture with BGP-LS

4. Including inter-domain connectivity in BGP-LS

In order for the parent PCE to carry out the path computation tasks it needs the inter-domain topology between the child domain scenarios. This topology is learnt through IGP by each BGP-LS speaker. The Traffic Engineering extensions (OSPF-TE or ISIS-TE) allow IGP to carry link state information that can be used in optimizing techniques such as the PCE algorithms. However, the parent PCE does not require such TE information, but just connectivity between the domains. However, TE information within the domain could be disseminated to the parent PCE to reduce the queries to the child PCEs.

4.1. Mapping from OSPF-TE

Carrying TE information in OSPF is a well-known standardized feature [RFC3630]. This section explains how this information can be exported outside one IGP domain using BGP-LS. BGP-LS extends the BGP Update messages to advertise link-state topology thanks to the new BGP Network Layer Reachability Information (NLRI) and BGP-LS attribute.

The BGP NLRI carries the descriptors used to define the element in question (e.g. link or node) and the BGP-LS attribute carries the chosen parameters to characterize the described element. Information is codified using multiple TLV triplets just as the ones used in OSPF-TE making it easy to integrate. For the purpose of this document, we consider a scenario where there is an origin (router) with the correspondent IPv4, a destination with its IPv4 and a link having the following TE parameters: maximum BW, maximum reservable BW and unreserved BW.

4.1.1. Node Descriptors

In the OSPF packet, there are two fields that tell us the origin and destination node IDs. The origin IP is the Source OSPF Router ID in the OSPF header and this is mapped into the IGP Router ID subTLV inside the Local Node Descriptors field [I-D.draft-ietf-idr-ls-distribution]. The destination IP is found as the Link ID field in the MPLS LSA in OSPF. This is mapped into the correspondent IGP Router ID in the Remote Node Descriptors field [I-D.draft-ietf-idr-ls-distribution].

There are other subTLVs inside the Local/Remote Node Descriptors but they are not relevant for this document.

4.1.2. Link Descriptors

The only two TLVs in the Link Descriptors field to map from OSPF are the local and remote interface addresses. This information is mapped directly from the Local/Remote Interface address TLV carried in the MPLS LSA of OSPF into the Local/Remote Interface address subTLV of the Link Descriptors field.

The same procedure must be applied for unnumbered interfaces but utilizing the Link Local/Remote Identifiers TLV.

4.1.3. Mapping OSPF TE parameters into BGP-LS attribute

As mentioned before, these parameters are not required in the H-PCE scenario. They are just required to reduce the number of queries to the children PCEs. The parent PCE can use bandwidth information between two domains to request for some possible connections instead of all.

The BGP-LS attribute will be a set of TLV triplets carrying the desired TE parameters learnt by OSPF. Bandwidth parameters are used to illustrate the example but they are many more (like, available labels).

The BGP-LS attribute is mapped in the following way. The TLVs carried in the MPLS-TE LSA in OSPF are directly translated into the equivalent TLVs in BGP-LS. As such, the Unreserved BW TLV in OSPF is mapped into the Unreserved BW TLV in BGP-LS. The same happens with the Maximum BW TLV and the Maximum Reservable BW TLV.

4.2. Mapping from ISIS-TE

TBD

5. Manageability Considerations

TBD

6. Security Considerations

TBD

7. References

7.1. Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC3630] Katz, D., Kompella, K., and D. Yeung, "Traffic Engineering (TE) Extensions to OSPF Version 2", RFC 3630, September 2003.
- [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.

7.2. Informative References

- [I-D.draft-dugeon-pce-ted-reqs]
Dugeon, O., Meuric, J., Douville, R., Casellas, R., and O. Gonzalez de Dios, "Path Computation Element (PCE) Traffic Engineering Database (TED) Requirements", February 2014.
- [I-D.draft-ietf-idr-ls-distribution]
Gredler, H., Medved, J., Previdi, S., Farrel, A., and S. Ray, "North-Bound Distribution of Link-State and TE Information using BGP", November 2013.
- [I-D.draft-ietf-idr-te-pm-bgp]
Wu, Q., Wang, D., Previdi, S., Gredler, H., and S. Ray, "BGP attribute for North-Bound Distribution of Traffic Engineering (TE) performance Metrics", January 2014.
- [I-D.draft-ietf-pce-hierarchy-extensions]
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)", July 2013.
- [I-D.draft-ietf-pce-inter-area-as-applicability]
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", February 2013.

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

Jeff Tantsura
Ericsson
USA

Email: jeff.tantsura@ericsson.com

PCE Working Group
Internet-Draft
Intended status: Standards Track
Expires: December 29, 2014

I. Minei
E. Crabbe
Google, Inc.
S. Sivabalan
Cisco Systems, Inc.
H. Ananthakrishnan
Juniper Networks, Inc.
X. Zhang
Huawei Technologies
Y. Tanaka
NTT Communications Corporation
June 27, 2014

PCEP Extensions for establishing relationships between sets of LSPs
draft-minei-pce-association-group-00

Abstract

This document introduces a generic mechanism to create a grouping of LSPs in the context of stateful PCE. This grouping can then be used to define associations between sets of LSPs or between a set of LSPs and a set of attributes (such as configuration parameters or behaviors), and is equally applicable to the active and passive modes of stateful PCE.

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 December 29, 2014.

Copyright Notice

Copyright (c) 2014 IETF Trust and the persons identified as the document authors. All rights reserved.

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

Table of Contents

1. Introduction	2
2. Terminology	3
3. Architectural Overview	3
3.1. Motivation	3
3.2. Operation overview	3
4. LSP association groups	4
5. Using the LSP association group	4
6. IANA considerations	5
7. Security Considerations	5
8. Acknowledgements	5
9. References	6
9.1. Normative References	6
9.2. Informative References	6
Authors' Addresses	6

1. Introduction

[RFC5440] describes the Path Computation Element Protocol PCEP. PCEP enables the communication between a Path Computation Client (PCC) and a Path Control Element (PCE), or between PCE and PCE, for the purpose of 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] and focuses on a model where LSPs are configured on the PCC and control over them is delegated to the PCE. The model of operation where LSPs are initiated from the PCE is described in [I-D.ietf-pce-pce-initiated-lsp].

This document introduces a generic mechanism to create a grouping of LSPs. This grouping can then be used to define associations between sets of LSPs or between a set of LSPs and a set of attributes (such as configuration parameters or behaviors), and is equally applicable to the active and passive modes of stateful PCE.

2. Terminology

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

3. Architectural Overview

3.1. Motivation

Stateful PCE provides the ability to update existing LSPs and to instantiate new ones. To enable support for PCE-controlled make-before-break and for protection, there is a need to define associations between LSPs. For example, the association between the original and the reoptimized path in the make-before break scenario, or between the working and protection path in end-to-end protection. Another use for LSP grouping is for applying a common set of configuration parameters or behaviors to a set of LSPs. Rather than creating separate mechanisms for each use case, this draft defines a generic one.

3.2. Operation overview

LSPs are associated with other LSPs with which they interact by adding them to a common association group. Association groups as defined in this document are locally meaningful at the LSP head-end, and can only be applied to LSPs originating at that head end. Thus, the association identifiers are unique at each head end, but not necessarily across the network, and are owned and managed by the head end.

Multiple types of groups can exist, each with their own identifiers space. The definition of the different association types and their behaviors is outside the scope of this document. The establishment and removal of the association relationship can be done on a per LSP basis. There is support for removal of all LSPs from an association as well. An LSP may join multiple association groups, of different or of the same type.

4. LSP association groups

Association groups are owned by the PCC, but the PCE may request creation of an association group (for example before instantiating LSPs that belong to that group). Membership in an association group can be initiated by either the PCE or the PCC. Association groups and their memberships are defined using the Association object.

The Association Object is an optional object in the PCUpd, PCRpt and PCinit messages.

The format of the Association object is shown Figure 1:

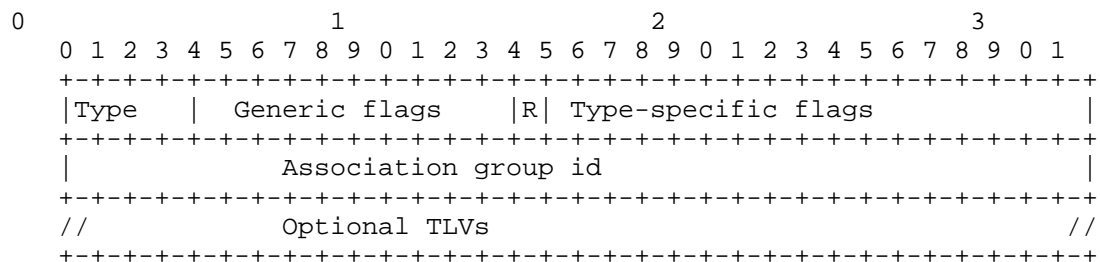


Figure 1: The Association Object format

Type - the association type (for example protection or make-before-break). The association type will be defined in separate documents.

Generic flags - flags for the association object. A single one is defined, the R flag indicating removal from the association group.

Type-specific flags - specific to the association type, will be defined at the time of the association type.

Association group id - identifier of the association group. The values 0 and 0xffffffff are reserved. Value 0 is used when the PCE requests allocation of an association group. Value 0xffffffff indicates all association groups.

5. Using the LSP association group

Membership in an association group is reported in PCRpt messages by including the association object along with the LSP object. Removal of the LSP from the association group on the PCC (for example through configuration) is reported by including the association object with the R flag set. When an LSP belongs to multiple association groups,

multiple association objects are included in the PCRpt, one for each association the LSP belongs to. A PCE can associate an LSP that was delegated to it (the candidate LSP) with an existing association group, by sending a PCUpd for the candidate LSP, including the Association Object for the association group. Error handling for this operation will be defined in a future version of this draft.

An association group can be created locally at the PCC (for example through configuration) or it can be requested by the PCE. A PCE may request the creation of an association group by sending a PCUpd message with the reserved value 0. In response to this request, the PCC will allocate an association group id and report it in the PCRpt message. Error handling will be defined in a future version of this draft. Note that this operation includes creation of the group and association of one LSP with this group. Requesting the creation of an association group before the LSP exists will be handled in a future version of this draft.

6. IANA considerations

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

Object-Class Value	Name	Reference
TBD	Association Object-Type 1	This document

This document requests that a registry is created to manage the Flags field of the Association object. New values are to be assigned by Standards Action [RFC5226].

7. 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, as the PCE may now create additional state on the PCC through the creation of association groups.

8. Acknowledgements

We would like to thank Yuji Kamite and Joshua George for their contributions to this document.

9. References

9.1. Normative References

- [I-D.ietf-pce-pce-initiated-lsp]
Crabbe, E., Minei, I., Sivabalan, S., and R. Varga, "PCEP Extensions for PCE-initiated LSP Setup in a Stateful PCE Model", draft-ietf-pce-pce-initiated-lsp-01 (work in progress), June 2014.
- [I-D.ietf-pce-stateful-pce]
Crabbe, E., Minei, I., Medved, J., and R. Varga, "PCEP Extensions for Stateful PCE", draft-ietf-pce-stateful-pce-09 (work in progress), June 2014.
- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC5226] Narten, T. and H. Alvestrand, "Guidelines for Writing an IANA Considerations Section in RFCs", BCP 26, RFC 5226, May 2008.
- [RFC5440] Vasseur, JP. and JL. Le Roux, "Path Computation Element (PCE) Communication Protocol (PCEP)", RFC 5440, March 2009.

9.2. Informative References

- [I-D.tanaka-pce-stateful-pce-mbb]
Tanaka, Y., Kamite, Y., and D. Dhody, "Make-Before-Break MPLS-TE LSP restoration and reoptimization procedure using Stateful PCE", draft-tanaka-pce-stateful-pce-mbb-03 (work in progress), February 2014.
- [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.

Authors' Addresses

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

Email: inaminei@google.com

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

Email: edc@google.com

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

Email: msiva@cisco.com

Hariharan Ananthakrishnan
Juniper Networks, Inc.
1194 N. Mathilda Ave.
Sunnyvale, CA 94089
US

Email: hanantha@juniper.net

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

Email: zhang.xian@huawei.com

Yosuke Tanaka
NTT Communications Corporation
Granpark Tower 3-4-1 Shibaura, Minato-ku
Tokyo 108-8118
Japan

Email: yosuke.tanaka@ntt.com

PCE Working Group
Internet-Draft
Intended status: Standards Track
Expires: January 5, 2015

U. Palle
D. Dhody
Huawei Technologies
Y. Tanaka
Y. Kamite
NTT Communications
Z. Ali
Cisco Systems
July 4, 2014

PCEP Extensions for PCE-initiated Point-to-Multipoint LSP Setup in a
Stateful PCE Model
draft-palle-pce-stateful-pce-initiated-p2mp-lsp-03

Abstract

The Path Computation Element (PCE) has been identified as an appropriate technology for the determination of the paths of point-to-multipoint (P2MP) TE LSPs. 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 PCE communication Protocol (PCEP), for a model where the Path Computation Client (PCC) delegates control over one or more locally configured LSPs to the PCE. Further [I-D.ietf-pce-pce-initiated-lsp] describes the creation and deletion of PCE-initiated LSPs under the stateful PCE model. This document provides extensions required for PCEP so as to enable the usage of a stateful PCE initiation capability in recommending point-to-multipoint (P2MP) TE LSP instantiation.

Status of This Memo

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

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

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

This Internet-Draft will expire on January 5, 2015.

Copyright Notice

Copyright (c) 2014 IETF Trust and the persons identified as the document authors. All rights reserved.

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

Table of Contents

1. Introduction	3
1.1. Requirements Language	3
2. Terminology	4
3. Architectural Overview	4
3.1. Motivation	4
3.2. Operation Overview	4
4. Support of PCE Initiated P2MP TE LSPs	5
5. PCE-initiated P2MP TE LSP Operations	5
5.1. The PCInitiate message	5
5.2. P2MP TE LSP Instantiation	6
5.3. P2MP TE LSP Deletion	7
5.4. Adding and Pruning Leaves for the P2MP TE LSP	7
5.5. P2MP TE LSP Delegation and Cleanup	7
6. PCInitiate Message Fragmentation	7
6.1. PCInitiate Fragmentation Procedure	8
7. Non-Support of P2MP TE LSP Instantiation for Stateful PCE	8
8. Security Considerations	8
9. Manageability Considerations	8
9.1. Control of Function and Policy	9
9.2. Information and Data Models	9
9.3. Liveness Detection and Monitoring	9
9.4. Verify Correct Operations	9
9.5. Requirements On Other Protocols	9
9.6. Impact On Network Operations	9
10. IANA Considerations	9
10.1. STATEFUL-PCE-CAPABILITY TLV	9
11. Acknowledgments	10
12. References	10
12.1. Normative References	10
12.2. Informative References	10

1. Introduction

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

[RFC4857] describes how to set up point-to-multipoint (P2MP) Traffic Engineering Label Switched Paths (TE LSPs) for use in Multiprotocol Label Switching (MPLS) and Generalized MPLS (GMPLS) networks. The PCE has been identified as a suitable application for the computation of paths for P2MP TE LSPs ([RFC5671]).

The PCEP is designed as a communication protocol between PCCs and PCEs for point-to-point (P2P) path computations and is defined in [RFC5440]. The extensions of PCEP to request path computation for P2MP TE LSPs are described in [RFC6006].

Stateful PCEs are shown to be helpful in many application scenarios, in both MPLS and GMPLS networks, as illustrated in [I-D.ietf-pce-stateful-pce-app]. These scenarios apply equally to P2P and P2MP TE LSPs. [I-D.ietf-pce-stateful-pce] provides the fundamental extensions needed for stateful PCE to support general functionality for P2P TE LSP. Further [I-D.palle-pce-stateful-pce-p2mp] focuses on the extensions that are necessary in order for the deployment of stateful PCEs to support P2MP TE LSPs. It includes mechanisms to effect P2MP LSP state synchronization between PCCs and PCEs, delegation of control of P2MP LSPs to PCEs, and PCE control of timing and sequence of P2MP path computations within and across PCEP sessions and focuses on a model where P2MP LSPs are configured on the PCC and control over them is delegated to the PCE.

[I-D.ietf-pce-pce-initiated-lsp] provides the fundamental extensions needed for stateful PCE-initiated P2P TE LSP recommended instantiation.

This document describes the setup, maintenance and teardown of PCE-initiated P2MP 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.

1.1. Requirements Language

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

2. Terminology

Terminology used in this document is same as terminology used in [I-D.ietf-pce-stateful-pce], [I-D.ietf-pce-pce-initiated-lsp] and [RFC6006].

3. Architectural Overview

3.1. Motivation

[I-D.palle-pce-stateful-pce-p2mp] provides stateful control over P2MP TE LSPs that are locally configured on the PCC. This model relies on the Ingress taking an active role in delegating locally configured P2MP TE LSPs to the PCE, and is well suited in environments where the P2MP TE LSP placement is fairly static. However, in environments where the P2MP TE LSP placement needs to change in response to application demands, it is useful to support dynamic creation and tear down of P2MP TE LSPs. The ability for a PCE to trigger the creation of P2MP TE LSPs on demand can be seamlessly integrated into a controller-based network architecture, where intelligence in the controller can determine when and where to set up paths.

Section 3 of [I-D.ietf-pce-pce-initiated-lsp] further describes the motivation behind the PCE-Initiation capability, which are equally applicable for P2MP TE LSPs.

3.2. Operation Overview

A PCC or PCE indicates its ability to support PCE provisioned dynamic P2MP LSPs during the PCEP Initialization Phase via mechanism described in Section 4.

As per section 5.1 of [I-D.ietf-pce-pce-initiated-lsp], the PCE sends a Path Computation LSP Initiate Request (PCInitiate) message to the PCC to suggest instantiation or deletion of a P2P TE LSP. This document extends the PCInitiate message to support P2MP TE LSP (see details in Section 5.1).

P2MP TE LSP suggested instantiation and deletion operations are same as P2P LSP as described in section 5.3 and 5.4 of [I-D.ietf-pce-pce-initiated-lsp]. This document focuses on extensions needed for further handling of P2MP TE LSP (see details in Section 5.2).

4. Support of PCE Initiated P2MP TE LSPs

During PCEP Initialization Phase, as per Section 7.1.1 of [I-D.ietf-pce-stateful-pce], PCEP speakers advertises Stateful capability via Stateful PCE Capability TLV in open message. A new flag is defined for the STATEFUL-PCE-CAPABILITY TLV defined in [I-D.ietf-pce-stateful-pce]. Its format is shown in the following figure:

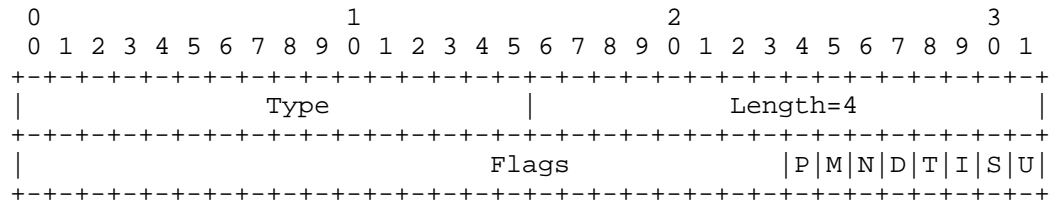


Figure 1: STATEFUL-PCE-CAPABILITY TLV Format

The U (LSP-UPDATE-CAPABILITY) bit is defined in [I-D.ietf-pce-stateful-pce]. The I (LSP-INSTITUTION-CAPABILITY) bit is defined in [I-D.ietf-pce-pce-initiated-lsp]. The S (INCLUDE-DB-VERSION), T (TRIGGERED-SYNC) and D (DELTA-LSP-SYNC-CAPABILITY) bits are defined in [I-D.ietf-pce-stateful-sync-optimizations]. The N (P2MP-CAPABILITY) and M (P2MP-LSP-UPDATE-CAPABILITY) bits are defined in [I-D.palle-pce-stateful-pce-p2mp]. A new bit P (P2MP-LSP-INSTITUTION-CAPABILITY) is added in this document:

P (P2MP-LSP-INSTITUTION-CAPABILITY - 1 bit): If set to 1 by a PCC, the P Flag indicates that the PCC allows suggested instantiation of an P2MP LSP by a PCE. If set to 1 by a PCE, the P flag indicates that the PCE will suggest P2MP LSP instantiation. The P2MP-LSP-INSTITUTION-CAPABILITY flag must be set by both PCC and PCE in order to support PCE-initiated P2MP LSP instantiation.

A PCEP speaker should continue to advertise the basic P2MP capability via mechanisms as described in [RFC6006].

5. PCE-initiated P2MP TE LSP Operations

5.1. The PCInitiate message

As defined in section 5.1 of [I-D.ietf-pce-pce-initiated-lsp], PCE sends a PCInitiate message to a PCC to recommend instantiation of a P2P TE LSP, this document extends the format of PCInitiate message for the creation of P2MP TE LSPs but the creation and deletion operations of P2MP TE LSP are same to the P2P TE LSP.

The format of PCInitiate message is as follows:

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

Where:

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

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

```
<PCE-initiated-lsp-instantiation> ::= <SRP>
                                       <LSP>
                                       <end-point-path-pair-list>
                                       [<attribute-list>]
```

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

Where:

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

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

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

The PCInitiate message with an LSP object with N bit (P2MP) set is used to convey operation on a P2MP TE LSP. 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 as described in [I-D.ietf-pce-stateful-pce].

5.2. P2MP TE LSP Instantiation

The Instantiation operation of P2MP TE LSP is same as defined in section 5.3 of [I-D.ietf-pce-pce-initiated-lsp] including handling of PLSP-ID, SYMBOLIC-PATH-NAME etc. Rules of processing and error codes remains unchanged. Further, as defined in section 6.1 of [I-D.palle-pce-stateful-pce-p2mp], N bit MUST be set in LSP object in

PCInitiate message by PCE to specify the instantiation is for P2MP TE LSP and the PCC or PCE MUST follow the mechanism defined in [I-D.palle-pce-stateful-pce-p2mp] for delegation and updation of P2MP TE LSPs.

Though N bit is set in the LSP object, P2MP-LSP-IDENTIFIER TLV defined in section 6.2 of [I-D.palle-pce-stateful-pce-p2mp] MUST NOT be included in the LSP object in PCInitiate message as it SHOULD be generated by PCC and carried in PCRpt message.

5.3. P2MP TE LSP Deletion

The deletion operation of P2MP TE LSP is same as defined in section 5.4 of [I-D.ietf-pce-pce-initiated-lsp] 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 (LSP-REMOVE as per section 5.2 of [I-D.ietf-pce-pce-initiated-lsp]). Rules of processing and error codes remains unchanged.

5.4. Adding and Pruning Leaves for the P2MP TE LSP

Adding of new leaves and Pruning of old Leaves for the PCE initiated P2MP TE LSP MUST be carried in PCUpd message and SHOULD refer [I-D.palle-pce-stateful-pce-p2mp] for P2MP TE LSP extensions. As defined in [RFC6006], leaf type = 1 for adding of new leaves, leaf type = 2 for pruning of old leaves of P2MP END-POINTS Object are used in PCUpd message.

PCC MAY use the Incremental State Update mechanisms as described in [RFC4875] to signal adding and pruning of leaves.

5.5. P2MP TE LSP Delegation and Cleanup

P2MP TE LSP delegation and cleanup operations are same as defined in section 6 of [I-D.ietf-pce-pce-initiated-lsp]. Rules of processing and error codes remains unchanged.

6. PCInitiate Message Fragmentation

The total PCEP message length, including the common header, is 16 bytes. In certain scenarios the P2MP LSP Initiate may not fit into a single PCEP message (initial PCInitiate message). The F-bit is used in the LSP object to signal that the initial PCInitiate was too large to fit into a single message and will be fragmented into multiple messages.

Fragmentation procedure described below for PCInitiate message is similar to [RFC6006] which describes request and response message fragmentation.

6.1. PCInitiate Fragmentation Procedure

Once the PCE initiates to set up the P2MP TE LSP, a PCInitiate message is sent to the PCC. If the PCInitiate is too large to fit into a single PCInitiate message, the PCE will split the PCInitiate over multiple messages. Each PCInitiate message sent by the PCE, except the last one, will have the F-bit set in the LSP object to signify that the PCInitiate has been fragmented into multiple messages. In order to identify that a series of PCInitiate messages represents a single Initiate, each message will use the same PLSP-ID (in this case 0) and SRP-ID-number.

[Editor Note: P2MP message fragmentation errors associated with a P2MP path initiation will be defined in future version].

7. Non-Support of P2MP TE LSP Instantiation for Stateful PCE

The PCEP protocol extensions described in this document for PCC or PCE with instantiation capability for P2MP TE LSPs MUST NOT be used if PCC or PCE has not advertised its stateful capability with Instantiation and P2MP capability as per Section 4. If this is not the case and Stateful initiation operations on P2MP TE LSPs are attempted, then a PCErr with error-type 19 (Invalid Operation) and error-value TBD needs to be generated.

[Editor Note: more information on exact error value is needed]

8. Security Considerations

The stateful operations on P2MP TE LSP are more CPU-intensive and also utilize more link bandwidth. In the event of an unauthorized stateful P2MP operations, or a denial of service attack, the subsequent PCEP operations may be disruptive to the network. Consequently, it is important that implementations conform to the relevant security requirements of [RFC5440], [RFC6006], [I-D.ietf-pce-stateful-pce] and [I-D.ietf-pce-pce-initiated-lsp].

9. Manageability Considerations

All manageability requirements and considerations listed in [RFC5440], [RFC6006], [I-D.ietf-pce-stateful-pce] and [I-D.ietf-pce-pce-initiated-lsp] apply to PCEP protocol extensions defined in this document. In addition, requirements and considerations listed in this section apply.

9.1. Control of Function and Policy

A PCE or PCC implementation **MUST** allow configuring the stateful Initiation capability for P2MP LSPs.

9.2. Information and Data Models

The PCEP MIB module **SHOULD** be extended to include advertised P2MP stateful PCE-Initiation capability etc.

9.3. Liveness Detection and Monitoring

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

9.4. Verify Correct Operations

Mechanisms defined in this document do not imply any new operation verification requirements in addition to those already listed in [RFC5440], [RFC6006] and [I-D.ietf-pce-stateful-pce].

9.5. Requirements On Other Protocols

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

9.6. Impact On Network Operations

Mechanisms defined in this document do not have any impact on network operations in addition to those already listed in [RFC5440], [RFC6006] and [I-D.ietf-pce-stateful-pce].

10. 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.

10.1. 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
25	P2MP-LSP- INSTANTIATION- CAPABILITY	This.I-D

11. Acknowledgments

Thanks to Quintin Zhao and Venugopal Reddy for his comments.

12. References

12.1. Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC5440] Vasseur, JP. and JL. Le Roux, "Path Computation Element (PCE) Communication Protocol (PCEP)", RFC 5440, March 2009.
- [I-D.ietf-pce-stateful-pce]
Crabbe, E., Minei, I., Medved, J., and R. Varga, "PCEP Extensions for Stateful PCE", draft-ietf-pce-stateful-pce-09 (work in progress), June 2014.
- [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-01 (work in progress), June 2014.
- [I-D.palle-pce-stateful-pce-p2mp]
Palle, U., Dhody, D., Tanaka, Y., Kamite, Y., and Z. Ali, "Path Computation Element (PCE) Protocol Extensions for Stateful PCE usage for Point-to-Multipoint Traffic Engineering Label Switched Paths", draft-palle-pce-stateful-pce-p2mp-03 (work in progress), June 2014.

12.2. Informative References

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

- [RFC4857] Fogelstroem, E., Jonsson, A., and C. Perkins, "Mobile IPv4 Regional Registration", RFC 4857, June 2007.
- [RFC4875] Aggarwal, R., Papadimitriou, D., and S. Yasukawa, "Extensions to Resource Reservation Protocol - Traffic Engineering (RSVP-TE) for Point-to-Multipoint TE Label Switched Paths (LSPs)", RFC 4875, May 2007.
- [RFC5671] Yasukawa, S. and A. Farrel, "Applicability of the Path Computation Element (PCE) to Point-to-Multipoint (P2MP) MPLS and GMPLS Traffic Engineering (TE)", RFC 5671, October 2009.
- [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.
- [I-D.ietf-pce-stateful-pce-app]
Zhang, X. and I. Minei, "Applicability of a Stateful Path Computation Element (PCE)", draft-ietf-pce-stateful-pce-app-02 (work in progress), June 2014.
- [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-01 (work in progress), June 2014.

Authors' Addresses

Udayasree Palle
Huawei Technologies
Leela Palace
Bangalore, Karnataka 560008
INDIA

EMail: udayasree.palle@huawei.com

Dhruv Dhody
Huawei Technologies
Leela Palace
Bangalore, Karnataka 560008
INDIA

EMail: dhruv.ietf@gmail.com

Yosuke Tanaka
NTT Communications Corporation
Granpark Tower
3-4-1 Shibaura, Minato-ku
Tokyo 108-8118
Japan

EMail: yosuke.tanaka@ntt.com

Yuji Kamite
NTT Communications Corporation
Granpark Tower
3-4-1 Shibaura, Minato-ku
Tokyo 108-8118
Japan

EMail: y.kamite@ntt.com

Zafar Ali
Cisco Systems

EMail: zali@cisco.com

PCE Working Group
Internet-Draft
Intended status: Standards Track
Expires: January 5, 2015

U. Palle
D. Dhody
Huawei Technologies
Y. Tanaka
Y. Kamite
NTT Communications
Z. Ali
Cisco Systems
July 4, 2014

Path Computation Element (PCE) Protocol Extensions for Stateful PCE
usage for Point-to-Multipoint Traffic Engineering Label Switched Paths
draft-palle-pce-stateful-pce-p2mp-04

Abstract

The Path Computation Element (PCE) has been identified as an appropriate technology for the determination of the paths of point-to-multipoint (P2MP) TE LSPs. [I-D.ietf-pce-stateful-pce-app] presents several use cases, demonstrating scenarios that benefit from the deployment of a stateful PCE. [I-D.ietf-pce-stateful-pce] provides the fundamental PCE communication Protocol (PCEP) extensions needed to support stateful PCE functions. This memo provides extensions required for PCEP so as to enable the usage of a stateful PCE capability in supporting point-to-multipoint (P2MP) TE LSPs.

Status of This Memo

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

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

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

This Internet-Draft will expire on January 5, 2015.

Copyright Notice

Copyright (c) 2014 IETF Trust and the persons identified as the document authors. All rights reserved.

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

Table of Contents

1. Introduction	3
1.1. Requirements Language	4
2. Terminology	4
3. Supporting P2MP TE LSP for Stateful PCE	4
3.1. Motivation	4
3.2. Objectives	4
4. Functions to Support P2MP TE LSPs for Stateful PCEs	4
5. Architectural Overview of Protocol Extensions	5
5.1. Extension of PCEP Messages	5
5.2. Capability Advertisement	6
5.3. State Synchronization	7
5.4. LSP Delegation	7
5.5. LSP Operations	7
5.5.1. Passive Stateful PCE	7
5.5.2. Active Stateful PCE	7
6. PCEP Object Extensions	7
6.1. Extension of LSP Object	7
6.2. P2MP-LSP-IDENTIFIER TLV	8
6.3. S2LS Object	11
7. PCEP Message Extensions	11
7.1. The PCRpt Message	11
7.2. The PCUpd Message	13
7.3. The PCReq Message	14
7.4. The PCRep Message	14
7.5. Example	15
7.5.1. P2MP TE LSP Update Request	15
7.5.2. P2MP TE LSP Report	15
7.6. Report and Update Message Fragmentation	16
7.6.1. Report Fragmentation Procedure	17
7.6.2. Update Fragmentation Procedure	17
8. Non-Support of P2MP TE LSPs for Stateful PCE	17

9. Security Considerations	18
10. Manageability Considerations	18
10.1. Control of Function and Policy	18
10.2. Information and Data Models	18
10.3. Liveness Detection and Monitoring	18
10.4. Verify Correct Operations	18
10.5. Requirements On Other Protocols	19
10.6. Impact On Network Operations	19
11. IANA Considerations	19
11.1. STATEFUL-PCE-CAPABILITY TLV	19
11.2. Extension of LSP Object	19
11.3. Extension of PCEP-Error Object	20
11.4. PCEP TLV Type Indicators	20
12. Acknowledgments	20
13. References	20
13.1. Normative References	20
13.2. Informative References	21

1. Introduction

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

[RFC4857] describes how to set up point-to-multipoint (P2MP) Traffic Engineering Label Switched Paths (TE LSPs) for use in Multiprotocol Label Switching (MPLS) and Generalized MPLS (GMPLS) networks. The PCE has been identified as a suitable application for the computation of paths for P2MP TE LSPs ([RFC5671]).

The PCEP is designed as a communication protocol between PCCs and PCEs for point-to-point (P2P) path computations and is defined in [RFC5440]. The extensions of PCEP to request path computation for P2MP TE LSPs are described in [RFC6006].

Stateful PCEs are shown to be helpful in many application scenarios, in both MPLS and GMPLS networks, as illustrated in [I-D.ietf-pce-stateful-pce-app]. These scenarios apply equally to P2P and P2MP TE LSPs. [I-D.ietf-pce-stateful-pce] provides the fundamental extensions needed for stateful PCE to support general functionality for P2P TE LSP. Complementarily, this document focuses on the extensions that are necessary in order for the deployment of stateful PCEs to support P2MP TE LSPs.

1.1. Requirements Language

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

2. Terminology

Terminology used in this document is same as terminology used in [I-D.ietf-pce-stateful-pce] and [RFC6006].

3. Supporting P2MP TE LSP for Stateful PCE

3.1. Motivation

[I-D.ietf-pce-stateful-pce-app] presents several use cases, demonstrating scenarios that benefit from the deployment of a stateful PCE including optimization, recovery, etc which are equally applicable to P2MP TE LSPs. [I-D.ietf-pce-stateful-pce] defines the extensions to PCEP for P2P TE LSPs. Complementarily, this document focuses on the extensions that are necessary in order for the deployment of stateful PCEs to support P2MP TE LSPs.

In addition to that, the stateful nature of a PCE simplifies the information conveyed in PCEP messages since it is possible to refer to the LSPs via PLSP-ID. For P2MP this is an added advantage, where the size of message is much larger. Incase of stateless PCE, a modification of 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 P2MP LSP identifier).

3.2. Objectives

The objectives for the protocol extensions to support P2MP TE LSP for stateful PCE are same as the objectives described in section 3.2 of [I-D.ietf-pce-stateful-pce].

4. Functions to Support P2MP TE LSPs for Stateful PCEs

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

This document extends these functions to support P2MP TE LSPs.

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 for P2MP using mechanisms defined in Section 5.2.

LSP State Synchronization (C-E): after the session between the PCC and a stateful PCE with P2MP capability is initialized, the PCE must learn the state of a PCC's P2MP TE LSPs before it can perform path computations or update LSP attributes in a PCC.

LSP Update Request (E-C): a stateful PCE with P2MP capability requests modification of attributes on a PCC's P2MP TE LSP.

LSP State Report (C-E): a PCC sends an LSP state report to a PCE whenever the state of a P2MP TE 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 P2MP TE LSPs; the PCE becomes the authoritative source of the LSP's attributes as long as the delegation is in effect (See Section 5.5 of [I-D.ietf-pce-stateful-pce]); the PCC may withdraw the delegation or the PCE may give up the delegation at any time.

An update to [I-D.sivabalan-pce-disco-stateful] is needed to support autodiscovery of stateful PCEs with P2MP capability.

5. Architectural Overview of Protocol Extensions

5.1. Extension of PCEP Messages

New PCEP messages are defined in [I-D.ietf-pce-stateful-pce] to support stateful PCE for P2P TE LSPs. In this document these messages are extended to support P2MP TE LSPs.

Path Computation State Report (PCRpt): Each P2MP TE LSP State Report in a PCRpt message can contain actual P2MP TE LSP path attributes, LSP status, etc. An LSP State Report carried on a PCRpt message is also used in delegation or revocation of control of a P2MP TE LSP to/from a PCE. The extension of PCRpt message is described in Section 7.1.

Path Computation Update Request (PCUpd): Each P2MP TE LSP Update Request in a PCUpd message MUST contain all LSP parameters that a PCE wishes to set for a given P2MP TE 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 a P2MP TE LSP. The PCUpd message is described in Section 7.2.

5.2. Capability Advertisement

During PCEP Initialization Phase, as per Section 7.1.1 of [I-D.ietf-pce-stateful-pce], PCEP speakers advertises Stateful capability via Stateful PCE Capability TLV in open message. A new flag is defined for the STATEFUL-PCE-CAPABILITY TLV defined in [I-D.ietf-pce-stateful-pce]. Its format is shown in the following figure:

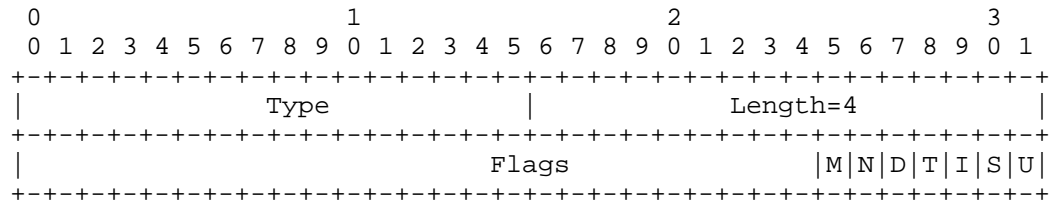


Figure 1: STATEFUL-PCE-CAPABILITY TLV Format

The U (LSP-UPDATE-CAPABILITY) bit is defined in [I-D.ietf-pce-stateful-pce]. The I (LSP-INSTITUTION-CAPABILITY) bit is defined in [I-D.ietf-pce-pce-initiated-lsp]. The S (INCLUDE-DB-VERSION), T (TRIGGERED-SYNC) and D (DELTA-LSP-SYNC-CAPABILITY) bits are defined in [I-D.ietf-pce-stateful-sync-optimizations]. A new bit N (P2MP-CAPABILITY) and M (P2MP-LSP-UPDATE-CAPABILITY) are added in this document:

N (P2MP-CAPABILITY - 1 bit): if set to 1 by a PCC, the N Flag indicates that the PCC is willing to send P2MP LSP State Reports whenever P2MP LSP parameters or operational status changes.; if set to 1 by a PCE, the N Flag indicates that the PCE is interested in receiving LSP State Reports whenever LSP parameters or operational status changes. The P2MP-CAPABILITY Flag must be advertised by both a PCC and a PCE for PCRpt messages P2MP extension to be allowed on a PCEP session.

M (P2MP-LSP-UPDATE-CAPABILITY - 1 bit): if set to 1 by a PCC, the M Flag indicates that the PCC allows modification of P2MP LSP parameters; if set to 1 by a PCE, the M Flag indicates that the PCE is capable of updating P2MP LSP parameters. The P2MP-LSP-UPDATE-CAPABILITY Flag must be advertised by both a PCC and a PCE for PCUpd messages P2MP extension to be allowed on a PCEP session.

A PCEP speaker should continue to advertise the basic P2MP capability via mechanisms as described in [RFC6006].

5.3. State Synchronization

State Synchronization operations described in Section 5.4 of [I-D.ietf-pce-stateful-pce] are applicable for P2MP TE LSPs as well.

5.4. LSP Delegation

LSP delegation operations described in Section 5.5 of [I-D.ietf-pce-stateful-pce] are applicable for P2MP TE LSPs as well.

5.5. LSP Operations

5.5.1. Passive Stateful PCE

LSP operations for passive stateful PCE described in Section 5.6.1 of [I-D.ietf-pce-stateful-pce] are applicable for P2MP TE LSPs as well.

The Path Computation Request and Response message format for P2MP TE LSPs is described in Section 3.4 and Section 3.5 of [RFC6006] respectively.

The Request and Response message for P2MP TE LSPs are extended to support encoding of LSP object, so that it is possible to refer to a LSP with a unique identifier and simplify the PCEP message exchange. For example, incase of modification of one leaf in a P2MP tree, there should be no need to carry the full P2MP tree in PCReq message.

The extension for the Request and Response message for passive stateful operations on P2MP TE LSPs are described in Section 7.3 and Section 7.4.

5.5.2. Active Stateful PCE

LSP operations for active stateful PCE described in Section 5.6.2 of [I-D.ietf-pce-stateful-pce] are applicable for P2MP TE LSPs as well.

6. PCEP Object Extensions

The PCEP TLV defined in this document is compliant with the PCEP TLV format defined in [RFC5440].

6.1. Extension of LSP Object

LSP Object is defined in Section 7.3 of [I-D.ietf-pce-stateful-pce]. It specifies PLSP-ID to uniquely identify an LSP that is constant for the life time of a PCEP session. Similarly for P2MP tunnel, PLSP-ID identify a P2MP TE LSP uniquely. This document adds the following flags to the LSP Object:

N (P2MP bit): If the bit is set to 1, it specifies the message is for P2MP TE LSP which MUST be set in PCRpt or PCUpd message for a P2MP TE LSP.

F (Fragmentation bit): If the bit is set to 1, it specifies the message is fragmented.

If P2MP bit is set, the following P2MP-LSP-IDENTIFIER TLV MUST be present in LSP object.

6.2. P2MP-LSP-IDENTIFIER TLV

The P2MP LSP Identifier TLV MUST be included in the LSP object in PCRpt message for RSVP-signaled P2MP TE LSPs. If the TLV is missing, the PCE will generate an error with error-type 6 (mandatory object missing) and error-value TBD (12) (P2MP-LSP-IDENTIFIERS TLV missing) and close the PCEP session.

The P2MP LSP Identifier TLV MAY be included in the LSP object in PCUpd message for RSVP-signaled P2MP TE 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 P2MP LSP Identifier TLVs, one for IPv4 and one for IPv6.

The format of the IPV4-P2MP-LSP-IDENTIFIER TLV is shown in the following figure:

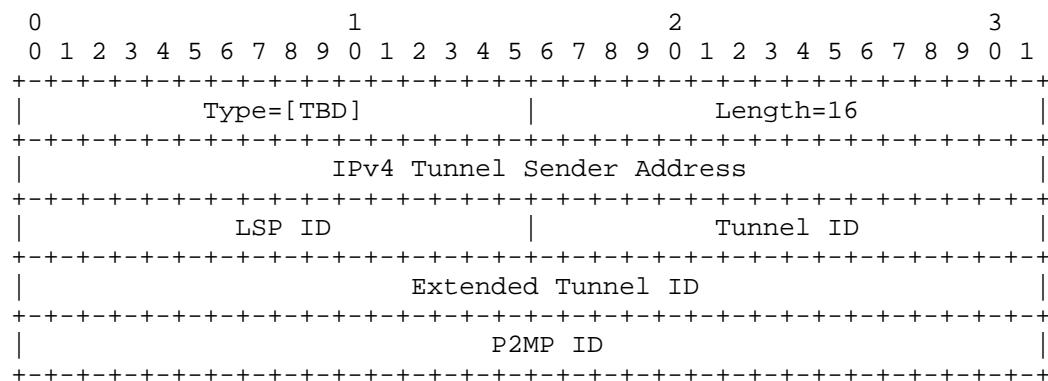


Figure 2: IPV4-P2MP-LSP-IDENTIFIER TLV format

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

Tunnel ID: contains the 16-bit 'Tunnel ID' identifier defined in [RFC3209], Section 4.6.1.1 for the LSP_TUNNEL_IPv4 Session Object. Tunnel ID remains constant over the life time of a tunnel.

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.

P2MP ID: contains the 32-bit 'P2MP ID' identifier defined in Section 19.1.1 of [RFC4875] for the P2MP LSP Tunnel IPv4 SESSION Object.

The format of the IPV6-P2MP-LSP-IDENTIFIER TLV is shown in the following figure:

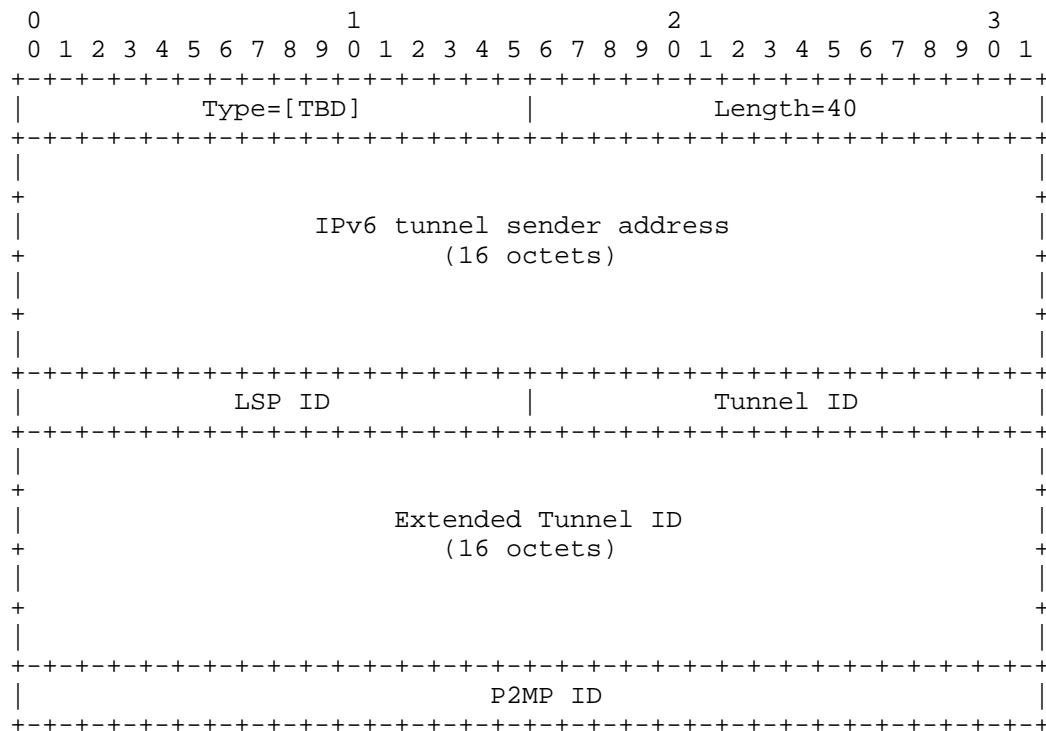


Figure 3: IPV6-P2MP-LSP-IDENTIFIER TLV format

The type of the TLV is [TBD] and it has a fixed length of 24 octets. 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.

Tunnel ID: contains the 16-bit 'Tunnel ID' identifier defined in [RFC3209], Section 4.6.1.2 for the LSP_TUNNEL_IPv6 Session Object. Tunnel ID remains constant over the life time of a tunnel.

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.

P2MP ID: As defined above in IPV4-P2MP-LSP-IDENTIFIERS TLV.

6.3. S2LS Object

The S2LS (Source-to-Leaves) Object is used to report RSVP state of one or more destinations (leaves) encoded within the END-POINTS object for a P2MP TE LSP. It MUST be carried in PCRpt message along with END-POINTS object when N bit is set in LSP object.

The format of the S2LS object is shown in the following figure:

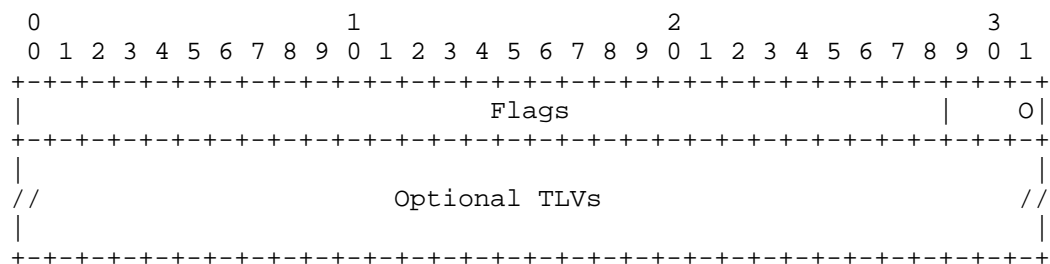


Figure 4: S2LS object format

Flags(32 bits):

O(Operational - 3 bits) the O Field represents the operational status of the group of destinations. The values are as per Operational field in LSP object defined in Section 7.3 of [I-D.ietf-pce-stateful-pce].

When N bit is set in LSP object then the O field in LSP object represents the operational status of the full P2MP TE LSP and the O field in S2LS object represents the operational status of a group of destinations encoded within the END-POINTS object.

Optional TLVs that may be included in the S2LS Object.

7. PCEP Message Extensions

7.1. The PCRpt Message

As per Section 6.1 of [I-D.ietf-pce-stateful-pce], PCRpt message is used to report the current state of a P2P TE LSP. This document extends the PCRpt message in reporting the status of P2MP TE LSP.

The format of 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>
                  <end-point-path-pair-list>
                  <attribute-list>
```

Where:

```
<end-point-path-pair-list> ::=
                        [<END-POINTS>]
                        [<S2LS>]
                        <path>
                        [<end-point-path-pair-list>]
```

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

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

The P2MP END-POINTS object defined in [RFC6006] is mandatory for specifying address of P2MP leaves grouped based on leaf types.

- o New leaves to add (leaf type = 1)
- o Old leaves to remove (leaf type = 2)
- o Old leaves whose path can be modified/reoptimized (leaf type = 3)
- o Old leaves whose path must be left unchanged (leaf type = 4)

When reporting the status of a P2MP TE LSP, the destinations are grouped in END-POINTS object based on the operational status (O field in S2LS object) and leaf type (in END-POINTS). This way the leaves that share the same operational status are grouped together. For reporting the status of delegated P2MP TE LSP, leaf-type = 3, where as for non-delegated P2MP TE LSP, leaf-type = 4 is used.

For delegated P2MP TE LSP configuration changes are reported via PCRpt message. For example, adding of new leaves END-POINTS (leaf-

type = 1) is used where as removing of old leaves (leaf-type = 2) is used.

Note that we preserve compatibility with the [I-D.ietf-pce-stateful-pce] definition of <state-report>. At least one instance of <END-POINTS> MUST be present in this message.

[Editor Note: suggest to add <END-POINTS> object mandatory in [I-D.ietf-pce-stateful-pce] document for <state-report>].

During state synchronization, the PCRpt message must report the status of the full P2MP TE LSP.

7.2. The PCUpd Message

As per Section 6.2 of [I-D.ietf-pce-stateful-pce], PCUpd message is used to update P2P TE LSP attributes. This document extends the PCUpd message in updating the attributes of P2MP TE LSP.

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>
                    <end-point-path-pair-list>
```

```
<attribute-list>
```

Where:

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

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

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

Note that we preserve compatibility with the [I-D.ietf-pce-stateful-pce] definition of <update-request>.

The PCC MAY use the make-before-break or sub-group-based procedures described in [RFC4875] based on a local policy decision.

7.3. The PCReq Message

As per Section 3.4 of [RFC6006], PCReq message is used for a P2MP path computation request. This document extends the PCReq message such that a PCC MAY include the LSP object in the PCReq message if the stateful PCE P2MP capability has been negotiated on a PCEP session between the PCC and a PCE.

The format of PCReq message is as follows:

```
<PCReq Message> ::= <Common Header>
                        <request>
```

where:

```
<request> ::= <RP>
              <end-point-rro-pair-list>
              [<LSP>]
              [<OF>]
              [<LSPA>]
              [<BANDWIDTH>]
              [<metric-list>]
              [<IRO>]
              [<LOAD-BALANCING>]
```

where:

```
<end-point-rro-pair-list> ::= <END-POINTS> [<RRO-List>] [<BANDWIDTH>]
                                [<end-point-rro-pair-list>]
```

```
<RRO-List> ::= <RRO> [<BANDWIDTH>] [<RRO-List>]
```

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

7.4. The PCRep Message

As per Section 3.5 of [RFC6006], PCRep message is used for a P2MP path computation reply. This document extends the PCRep message such that a PCE MAY include the LSP object in the PCRep message if the stateful PCE P2MP capability has been negotiated on a PCEP session between the PCC and a PCE.

The format of PCRep message is as follows:


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

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

where:

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

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

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

7.5. Example

7.5.1. P2MP TE LSP Update Request

LSP Update Request message is sent by an active stateful PCE to update the P2MP TE LSP parameters or attributes. An example of a PCUpd message for P2MP TE LSP is described below:

```
Common Header
SRP
LSP with P2MP flag set
END-POINTS for leaf type 3
ERO list
```

In this example, a stateful PCE request updation of path taken by some of the leaves in a P2MP tree. The update request uses the END-POINT type 3 (modified/reoptimized). The ERO list represents the S2LS path after modification. The update message does not need to encode the full P2MP tree in this case.

7.5.2. P2MP TE LSP Report

LSP State Report message is sent by a PCC to report or delegate the P2MP TE LSP. An example of a PCRpt message for a delegated P2MP TE LSP is described below to add new leaves to an existing P2MP TE LSP:

```
Common Header
LSP with P2MP flag set
END-POINTS for leaf type 1
  S2LS (O=DOWN)
  ERO list (empty)
```

An example of a PCRpt message for P2MP TE LSP is described below to prune leaves from an existing P2MP TE LSP:

```
Common Header
LSP with P2MP flag set
END-POINTS for leaf type 2
  S2LS (O=UP)
  ERO list
```

An example of a PCRpt message for a delegated P2MP TE LSP is described below to report status of leaves in an existing P2MP TE LSP:

```
Common Header
LSP with P2MP flag set
END-POINTS for leaf type 3
  S2LS (O=UP)
  ERO list
END-POINTS for leaf type 3
  S2LS (O=DOWN)
  ERO list
```

An example of a PCRpt message for a non-delegated P2MP TE LSP is described below to report status of leaves:

```
Common Header
LSP with P2MP flag set
END-POINTS for leaf type 4
  S2LS (O=ACTIVE)
  ERO list
END-POINTS for leaf type 4
  S2LS (O=DOWN)
  ERO list
```

7.6. Report and Update Message Fragmentation

The total PCEP message length, including the common header, is 16 bytes. In certain scenarios the P2MP report and update request may not fit into a single PCEP message (initial report or update). The

F-bit is used in the LSP object to signal that the initial report or update was too large to fit into a single message and will be fragmented into multiple messages. In order to identify the single report or update, each message will use the same PLSP-ID.

Fragmentation procedure described below for report or update message is similar to [RFC6006] which describes request and response message fragmentation.

7.6.1. Report Fragmentation Procedure

If the initial report is too large to fit into a single report message, the PCC will split the report over multiple messages. Each message sent to the PCE, except the last one, will have the F-bit set in the LSP object to signify that the report has been fragmented into multiple messages. In order to identify that a series of report messages represents a single report, each message will use the same PLSP-ID.

7.6.2. Update Fragmentation Procedure

Once the PCE computes and updates a path for some or all leaves in a P2MP TE LSP, an update message is sent to the PCC. If the update is too large to fit into a single update message, the PCE will split the update over multiple messages. Each update message sent by the PCE, except the last one, will have the F-bit set in the LSP object to signify that the update has been fragmented into multiple messages. In order to identify that a series of update messages represents a single update, each message will use the same PLSP-ID and SRP-ID-number.

[Editor Note: P2MP message fragmentation errors associated with a P2MP path report and update will be defined in future version].

8. Non-Support of P2MP TE LSPs for Stateful PCE

The PCEP protocol extensions described in this document for stateful PCEs with P2MP capability MUST NOT be used if PCE has not advertised its stateful capability with P2MP as per Section 5.2. If this is not the case and Stateful operations on P2MP TE LSPs are attempted, then a PCERR with error-type 19 (Invalid Operation) and error-value TBD needs to be generated.

If a Stateful PCE receives a P2MP TE LSP report message and it understands the P2MP flag in the LSP object, but the stateful PCE is not capable of P2MP computation, the PCE MUST send a PCERR message with error-type 19 (Invalid Operation) and error-value TBD.

If a Stateful PCE receives a P2MP TE LSP report message and the PCE does not understand the P2MP flag in the LSP object, and therefore the PCEP extensions described in this document, then the PCE SHOULD reject the request.

[Editor Note: more information on exact error value is needed]

9. Security Considerations

The stateful operations on P2MP TE LSP are more CPU-intensive and also utilize more link bandwidth. In the event of an unauthorized stateful P2MP operations, or a denial of service attack, the subsequent PCEP operations may be disruptive to the network. Consequently, it is important that implementations conform to the relevant security requirements of [RFC5440], [RFC6006] and [I-D.ietf-pce-stateful-pce].

10. Manageability Considerations

All manageability requirements and considerations listed in [RFC5440], [RFC6006] and [I-D.ietf-pce-stateful-pce] apply to PCEP protocol extensions defined in this document. In addition, requirements and considerations listed in this section apply.

10.1. Control of Function and Policy

A PCE or PCC implementation MUST allow configuring the stateful PCEP capability and the LSP Update capability for P2MP LSPs.

10.2. Information and Data Models

The PCEP MIB module SHOULD be extended to include advertised P2MP stateful capabilities, P2MP synchronization status, and P2MP delegation status etc.

10.3. 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].

10.4. Verify Correct Operations

Mechanisms defined in this document do not imply any new operation verification requirements in addition to those already listed in [RFC5440], [RFC6006] and [I-D.ietf-pce-stateful-pce].

10.5. Requirements On Other Protocols

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

10.6. 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], [RFC6006] and [I-D.ietf-pce-stateful-pce].

11. 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.

11.1. 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
27	P2MP-CAPABILITY	This.I-D
26	P2MP-LSP-UPDATE-CAPABILITY	This.I-D

11.2. Extension of LSP Object

This document requests that a registry is created to manage the Flags field of the LSP object. New values are to be assigned by Standards Action [RFC5226]. Each bit should be tracked with the following qualities:

- 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
24	P2MP	This.I-D
23	Fragmentation	This.I-D

11.3. Extension of PCEP-Error Object

A new error types 6 and 19 defined in section 8.4 of [I-D.ietf-pce-stateful-pce]. This document extend the new Error-Values for those error types for the following error conditions:

Error-Type	Meaning
6	Mandatory Object missing Error-value=12: P2MP-LSP-IDENTIFIER TLV missing
19	Invalid Operation Error-value= TBD.

11.4. PCEP TLV Type Indicators

This document defines the following new PCEP TLVs:

Value	Meaning	Reference
22	P2MP-IPV4-LSP-IDENTIFIERS	This.I-D
23	P2MP-IPV6-LSP-IDENTIFIERS	This.I-D

12. Acknowledgments

Thanks to Quintin Zhao and Venugopal Reddy for his comments.

13. References

13.1. Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC5440] Vasseur, JP. and JL. Le Roux, "Path Computation Element (PCE) Communication Protocol (PCEP)", RFC 5440, March 2009.

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

Crabbe, E., Minei, I., Medved, J., and R. Varga, "PCEP Extensions for Stateful PCE", draft-ietf-pce-stateful-pce-09 (work in progress), June 2014.

13.2. 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.
- [RFC4655] Farrel, A., Vasseur, J., and J. Ash, "A Path Computation Element (PCE)-Based Architecture", RFC 4655, August 2006.
- [RFC4857] Fogelstroem, E., Jonsson, A., and C. Perkins, "Mobile IPv4 Regional Registration", RFC 4857, June 2007.
- [RFC4875] Aggarwal, R., Papadimitriou, D., and S. Yasukawa, "Extensions to Resource Reservation Protocol - Traffic Engineering (RSVP-TE) for Point-to-Multipoint TE Label Switched Paths (LSPs)", RFC 4875, May 2007.
- [RFC5226] Narten, T. and H. Alvestrand, "Guidelines for Writing an IANA Considerations Section in RFCs", BCP 26, RFC 5226, May 2008.
- [RFC5671] Yasukawa, S. and A. Farrel, "Applicability of the Path Computation Element (PCE) to Point-to-Multipoint (P2MP) MPLS and GMPLS Traffic Engineering (TE)", RFC 5671, October 2009.
- [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.
- [I-D.ietf-pce-stateful-pce-app]
- Zhang, X. and I. Minei, "Applicability of a Stateful Path Computation Element (PCE)", draft-ietf-pce-stateful-pce-app-02 (work in progress), June 2014.
- [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-01 (work in progress), June 2014.

[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-01 (work in
progress), June 2014.

[I-D.sivabalan-pce-disco-stateful]

Sivabalan, S., Medved, J., and X. Zhang, "IGP Extensions
for Stateful PCE Discovery", draft-sivabalan-pce-disco-
stateful-03 (work in progress), January 2014.

Authors' Addresses

Udayasree Palle
Huawei Technologies
Leela Palace
Bangalore, Karnataka 560008
INDIA

EMail: udayasree.palle@huawei.com

Dhruv Dhody
Huawei Technologies
Leela Palace
Bangalore, Karnataka 560008
INDIA

EMail: dhruv.ietf@gmail.com

Yosuke Tanaka
NTT Communications Corporation
Granpark Tower
3-4-1 Shibaura, Minato-ku
Tokyo 108-8118
Japan

EMail: yosuke.tanaka@ntt.com

Yuji Kamite
NTT Communications Corporation
Granpark Tower
3-4-1 Shibaura, Minato-ku
Tokyo 108-8118
Japan

EMail: y.kamite@ntt.com

Zafar Ali
Cisco Systems

EMail: zali@cisco.com

PCE Working Group
Internet-Draft
Intended status: Standards Track
Expires: January 1, 2015

S. Sivabalan
J. Medved
Cisco Systems, Inc.
I. Minei
E. Crabbe
Google, Inc.
R. Varga
Pantheon Technologies SRO
June 30, 2014

Conveying path setup type in PCEP messages
draft-sivabalan-pce-lsp-setup-type-02.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 January 1, 2015.

Copyright Notice

Copyright (c) 2014 IETF Trust and the persons identified as the document authors. All rights reserved.

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

Table of Contents

1. Introduction	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.ietf-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.

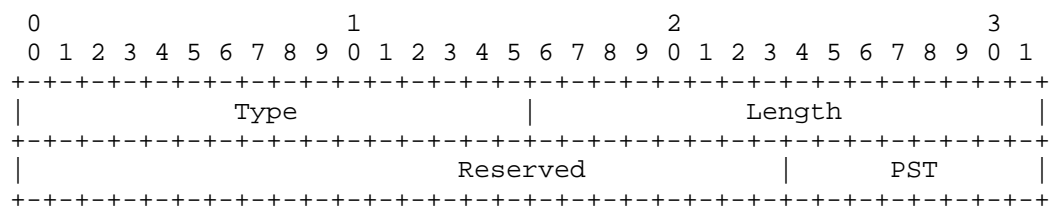


Figure 1: PATH-SETUP-TYPE TLV

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).

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

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.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-01 (work in progress), June 2014.
- [I-D.ietf-pce-stateful-pce]
Crabbe, E., Minei, I., Medved, J., and R. Varga, "PCEP Extensions for Stateful PCE", draft-ietf-pce-stateful-pce-09 (work in progress), June 2014.
- [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
Google, Inc.
1600 Amphitheatre Parkway
Mountain View, CA 94043
USA

Email: inaminei@google.com

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

Email: edc@google.com

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

Email: robert.vargad@pantheon.sk

Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: January 4, 2015

S. Sivabalan
J. Medved
C. Filsfils
Cisco Systems, Inc.
E. Crabbe
Google, Inc.
R. Raszuk
NTT I3
V. Lopez
Telefonica I+D
J. Tantsura
Ericsson
July 03, 2014

PCEP Extensions for Segment Routing
draft-sivabalan-pce-segment-routing-03.txt

Abstract

Segment Routing (SR) enables any head-end node to select any path without relying on a hop-by-hop signaling technique (e.g., LDP or RSVP-TE). It depends only on "segments" that are advertised by Link-State Interior Gateway Protocols (IGPs). A Segment Routed Path can be derived from a variety of mechanisms, including an IGP Shortest Path Tree (SPT), explicit configuration, or a Path Computation Element (PCE). This document specifies extensions to the Path Computation Element Protocol (PCEP) that allow a stateful PCE to compute and initiate Traffic Engineering (TE) paths, as well as a PCC to request a path subject to certain constraint(s) and optimization criteria in SR networks.

Requirements Language

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

Status of This Memo

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

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

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

This Internet-Draft will expire on January 4, 2015.

Copyright Notice

Copyright (c) 2014 IETF Trust and the persons identified as the document authors. All rights reserved.

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

Table of Contents

1. Introduction	3
2. Terminology	4
3. Overview of PCEP Operation in SR Networks	5
4. SR-Specific PCEP Message Extensions	6
5. Object Formats	7
5.1. The OPEN Object	7
5.1.1. The SR PCE Capability TLV	7
5.2. The RP/SRP Object	8
5.3. ERO Object	8
5.3.1. SR-ERO Subobject	9
5.3.2. NAI Associated with SID	10
5.3.3. ERO Processing	12
5.4. RRO Object	13
5.4.1. RRO Processing	13
6. Backward Compatibility	14
7. Management Considerations	14
7.1. Policy	14
7.2. The PCEP Data Model	14
8. Security Considerations	14
9. IANA Considerations	14
9.1. PCEP Objects	14
9.2. PCEP-Error Object	14
9.3. PCEP TLV Type Indicators	15
9.4. New Path Setup Type	15

10. Contributors	15
11. Acknowledgements	15
12. References	15
12.1. Normative References	15
12.2. Informative References	17
Authors' Addresses	17

1. Introduction

SR technology leverages the source routing and tunneling paradigms. A source node can choose a path without relying on hop-by-hop signaling protocols such as LDP or RSVP-TE. Each path is specified as a set of "segments" advertised by link-state routing protocols (IS-IS or OSPF). [I-D.filsfils-rtgwg-segment-routing] provides an introduction to SR architecture. The corresponding IS-IS and OSPF extensions are specified in [I-D.ietf-isis-segment-routing-extensions] and [I-D.ietf-ospf-segment-routing-extensions], respectively. SR architecture defines a "segment" as a piece of information advertised by a link-state routing protocols, e.g. an IGP prefix or an IGP adjacency. Several types of segments are defined. A Node segment represents an ECMP-aware shortest-path computed by IGP to a specific node, and is always global within SR/IGP domain. An Adjacency Segment represents unidirectional adjacency. An Adjacency Segment is local to the node which advertises it. Both Node segments and Adjacency segments can be used for SR Traffic Engineering (SR-TE).

The SR architecture can be applied to the MPLS forwarding plane without any change, in which case an SR path corresponds to an MPLS Label Switching Path (LSP). This document is relevant to only MPLS forwarding plane, and assumes that a 32-bit Segment Identifier (SID) represents an absolute value of MPLS label entry. In this document, "Node-SID" and "Adjacency-SID" denote Node Segment Identifier and Adjacency Segment Identifier respectively.

A Segment Routed path (SR path) can be derived from an IGP Shortest Path Tree (SPT). SR-TE paths may not follow IGP SPT. Such paths may be chosen by a suitable network planning tool and provisioned on the source node of the SR-TE path.

[RFC5440] describes Path Computation Element Protocol (PCEP) for communication between a Path Computation Client (PCC) and a Path Computation Element (PCE) or between one a pair of PCEs. A PCE or a PCC operating as a PCE (in hierarchical PCE environment) computes paths for MPLS Traffic Engineering LSPs (MPLS-TE LSPs) based on various constraints and optimization criteria. [I-D.ietf-pce-stateful-pce] specifies extensions to PCEP that allow a stateful PCE to compute and recommend network paths in compliance

with [RFC4657] and defines objects and TLVs for MPLS-TE LSPs. Stateful PCEP extensions provide synchronization of LSP state between a PCC and a PCE or between a pair of PCEs, delegation of LSP control, reporting of LSP state from a PCC to a PCE, controlling the setup and path routing of an LSP from a PCE to a PCC. Stateful PCEP extensions are intended for an operational model in which LSPs are configured on the PCC, and control over them is delegated to the PCE.

A mechanism to dynamically initiate LSPs on a PCC based on the requests from a stateful PCE or a controller using stateful PCE is specified in [I-D.ietf-pce-pce-initiated-lsp]. Such mechanism is useful in Software Driven Networks (SDN) applications, such as demand engineering, or bandwidth calendaring.

It is possible to use a stateful PCE for computing one or more SR-TE paths taking into account various constraints and objective functions. Once a path is chosen, the stateful PCE can initiate an SR-TE path on a PCC using PCEP extensions specified in [I-D.ietf-pce-pce-initiated-lsp] using the SR specific PCEP extensions described in this document. Additionally, using procedures described in this document, a PCC can request an SR path from either stateful or a stateless PCE. This specification relies on the PATH-SETUP-TYPE TLV and procedures specified in [I-D.sivabalan-pce-lsp-setup-type].

2. Terminology

The following terminologies are used in this document:

ERO: Explicit Route Object

IGP: Interior Gateway Protocol

IS-IS: Intermediate System to Intermediate System

LSR: Label Switching Router

MSD: Maximum SID Depth

NAI: Node or Adjacency Identifier

OSPF: Open Shortest Path First

PCC: Path Computation Client

PCE: Path Computation Element

PCEP: Path Computation Element Protocol

RRO: Record Route Object

SID: Segment Identifier

SR: Segment Routing

SR-TE: Segment Routed Traffic Engineering

TED: Traffic Engineering Database

3. Overview of PCEP Operation in SR Networks

In SR networks, an ingress node of an SR path appends all outgoing packets with an SR header consisting of a list of SIDs (or MPLS labels in the context of this document). The header has all necessary information to guide the packets from the ingress node to the egress node of the path, and hence there is no need for any signaling protocol.

In a PCEP session, LSP information is carried in the Explicit Route Object (ERO), which consists of a sequence of subobjects. Various types of ERO subobjects have been specified in [RFC3209], [RFC3473], and [RFC3477]. In SR networks, an ingress node of an SR path appends all outgoing packets with an SR header consisting of a list of SIDs (or MPLS labels in the context of this document). SR-TE LSPs computed by a PCE can be represented in one of the following forms:

- o An ordered set of IP address(es) representing network nodes/links: In this case, the PCC needs to convert the IP address(es) into the corresponding MPLS labels by consulting its Traffic Engineering Database (TED).
- o An ordered set of SID(s).
- o An ordered set of both MPLS label(s) and IP address(es): In this case, the PCC needs to convert the IP address(es) into the corresponding SID(s) by consulting its TED.

This document defines a new ERO subobject denoted by "SR-ERO subobject" capable of carrying a SID as well as the identity of the node/adjacency represented by the SID. SR-capable PCEP speakers should be able to generate and/or process such ERO subobject. An ERO containing SR-ERO subobjects can be included in the PCEP Path Computation Reply (PCRep) message defined in [RFC5440], the PCEP LSP Initiate Request message (PCInitiate) defined in [I-D.ietf-pce-pce-initiated-lsp], as well as in the PCEP LSP Update Request (PCUpd) and PCEP LSP State Report (PCRpt) messages defined in [I-D.ietf-pce-stateful-pce].

When a PCEP session between a PCC and a PCE is established, both PCEP speakers exchange information to indicate their ability to support SR-specific functionality. Furthermore, an LSP initially established via RSVP-TE signaling can be updated with SR-TE path. This capability is useful when a network is migrated from RSVP-TE to SR-TE technology. Similarly, an LSP initially created with SR-TE path can be updated to signal the LSP using RSVP-TE if necessary.

A PCC MAY include an RRO object containing the recorded LSP in PCReq and PCRpt messages as specified in [RFC5440] and [I-D.ietf-pce-stateful-pce] respectively. This document defines a new RRO subobject for SR networks. Methods used by a PCC to record SR-TE LSP are outside the scope of this document.

In summary, this document:

- o Defines a new PCEP capability, new ERO subobject, new RRO subobject, a new TLV, and new PCEP error codes.
- o Specifies how two PCEP speakers can establish a PCEP session that can carry information about SR-TE paths.
- o Specifies processing rules of ERO subobject.
- o Defines a new path setup type carried in the PATH-SETUP-TYPE TLV for SR-TE LSP.

The extensions specified in this document are applicable to the stateless PCE model defined in [RFC5440], as well as for the active stateful and passive stateful PCE models defined in [I-D.ietf-pce-stateful-pce].

4. SR-Specific PCEP Message Extensions

As defined in [RFC5440], a PCEP message consists of a common header followed by a variable length body made up of mandatory and/or optional objects. This document does not require any changes in the format of PCReq and PCRep messages specified in [RFC5440], PCInitiate message specified in [I-D.ietf-pce-pce-initiated-lsp], and PCRpt and PCUpd messages specified in [I-D.ietf-pce-stateful-pce]. However, PCEP messages pertaining to SR-TE LSP MUST include PATH-SETUP-TYPE TLV in the RP or SRP object to clearly identify that SR-TE LSP is intended. In other words, a PCEP speaker MUST not infer whether or not a PCEP message pertains to SR-TE LSP from any other object or TLV.

5. Object Formats

5.1. The OPEN Object

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

5.1.1. The SR PCE Capability TLV

The SR-PCE-CAPABILITY TLV is an optional TLV associated with the OPEN Object to exchange SR capability of PCEP speakers. The format of the SR-PCE-CAPABILITY TLV is shown in the following figure:

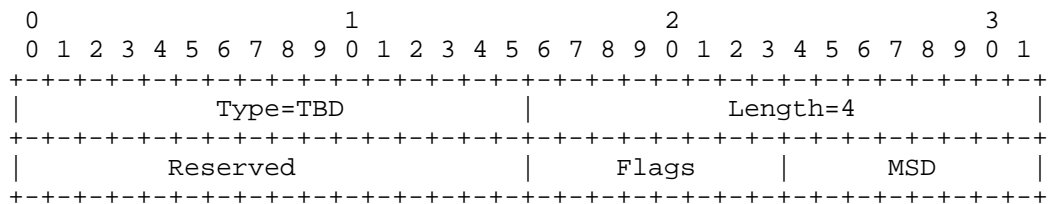


Figure 1: SR-PCE-CAPABILITY TLV format

The code point for the TLV type is to be defined by IANA. The TLV length is 4 octets.

The 32-bit value is formatted as follows. The "Maximum SID Depth" (1 octet) field (MSD) specifies the maximum number of SIDs that a PCC is capable of imposing on a packet. The "Flags" (1 octet) and "Reserved" (2 octets) fields are currently unused, and MUST be set to zero on transmission and ignored on reception.

5.1.1.1. Exchanging SR Capability

By including the SR-PCE-CAPABILITY TLV in the OPEN message destined to a PCE, a PCC indicates that it is capable of supporting the head-end functions for SR-TE LSP. By including the TLV in the OPEN message destined to a PCC, a PCE indicates that it is capable of computing SR-TE paths.

The number of SIDs that can be imposed on a packet depends on PCC's data plane's capability. The default value of MSD is 0 meaning that a PCC does not impose any limitation on the number of SIDs included in any SR-TE path coming from PCE. Once an SR-capable PCEP session is established with a non-default MSD value, the corresponding PCE cannot send SR-TE paths with SIDs exceeding that MSD value. If a PCC needs to modify the MSD value, the PCEP session MUST be closed and re-established with the new MSD value. If a PCEP session is

established with a non-default MSD value, and the PCC receives an SR-TE path containing more SIDs than specified in the MSD value, the PCC MUST send a PCErr message with Error-Type 10 (Reception of an invalid object) and Error-value 3 (Unsupported number of Segment ERO).

The SR Capability TLV is meaningful only in the OPEN message sent from a PCC to a PCE. As such, a PCE does not need to set MSD value in outbound message to a PCC. Similarly, a PCC ignores any MSD value received from a PCE. If a PCE receives multiple SR-PCE-CAPABILITY TLVs in an OPEN message, it processes only the first TLV is processed.

5.2. The RP/SRP Object

In order to setup an SR-TE LSP using SR, RP or SRP object MUST PATH-SETUP-TYPE TLV specified in [I-D.sivabalan-pce-lsp-setup-type]. This document defines a new Path Setup Type (PST) for SR as follows:

- o PST = 1: Path is setup using Segment Routing Traffic Engineering technique.

5.3. ERO Object

An SR-TE path consists of one or more SID(s) where each SID MAY be associated with the identifier that represents the node or adjacency corresponding to the SID. This identifier is referred to as the 'Node or Adjacency Identifier' (NAI). As described later, a NAI can be represented in various formats (e.g., IPv4 address, IPv6 address, etc). Furthermore, a NAI is used only for troubleshooting purposes, and MUST NOT be used to replace or modify any fields in a data packet header.

The ERO object specified in [RFC5440] is used to carry SR-TE path information. In order to carry SID and/or NAI, this document defines a new ERO subobject referred to as "SR-ERO subobject" whose format is specified in the following section. An ERO object carrying an SR-TE path consists of one or more ERO subobject(s), and MUST carry only SR-ERO subobject. Note that an SR-ERO subobject does not need to have both SID and NAI. However, at least one of them MUST be present.

When building the MPLS label stack from ERO, a PCC MUST assume that SR-ERO subobjects are organized as a last-in-first-out stack. The first subobject relative to the beginning of ERO contains the information about the topmost label. The last subobject contains information about the bottommost label.

5.3.1. SR-ERO Subobject

An SR-ERO subobject consists of a 32-bit header followed by the SID and the NAI associated with the SID. The SID is a 32-bit number. The size of the NAI depends on its respective type, as described in the following sections.

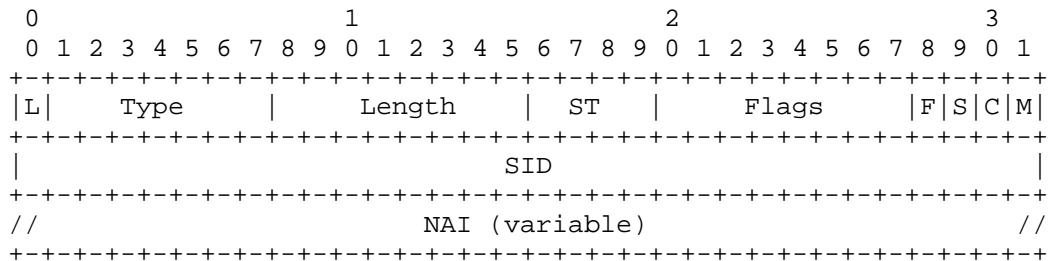


Figure 2: SR-ERO Subobject format

The fields in the SR-ERO Subobject are as follows:

The 'L' Flag indicates whether the subobject represents a loose-hop in the LSP [RFC3209]. If this flag is unset, a PCC MUST not overwrite the SID value present in the SR-ERO subobject. Otherwise, a PCC MAY expand or replace one or more SID value(s) in the received SR-ERO based on its local policy.

Type is the type of the SR-ERO subobject. This document defines the SR-ERO subobject type, and requests a new codepoint from IANA.

Length contains the total length of the subobject in octets, including the L, Type and Length fields. Length MUST be at least 8, and MUST be a multiple of 4. As mentioned earlier, an SR-ERO subobject MUST have at least SID or NAI. The length should take into consideration SID or NAI only if they are not null. The flags described below used to indicate whether SID or NAI field is null.

SID Type (ST) indicates the type of information associated with the SID contained in the object body. The SID-Type values are described later in this document.

Flags is used to carry any additional information pertaining to SID. Currently, the following flag bits are defined:

- * M: When this bit is set, the SID value represents an MPLS label stack entry as specified in [RFC5462] where only the label value is specified by the PCE. Other fields (TC, S, and TTL) fields MUST be considered invalid, and PCC MUST set these fields according to its local policy and MPLS forwarding rules.
- * C: When this bit as well as the M bit are set, then the SID value represents an MPLS label stack entry as specified in [RFC5462], where all the entry's fields (Label, TC, S, and TTL) are specified by the PCE. However, a PCC MAY choose to override TC, S, and TTL values according its local policy and MPLS forwarding rules.
- * S: When this bit is set, the SID value in the subobject body is null. In this case, the PCC is responsible for choosing the SID value, e.g., by looking up its TED using the NAI which, in this case, MUST be present in the subobject.
- * F: When this bit is set, the NAI value in the subobject body is null.

Editorial Note: we need to decide how to treat an SR-ERO subobject in which both NAI and SID are null.

SID is the Segment Identifier.

NAI contains the NAI associated with the SID. Depending on the value of ST, the NAI can have different format as described in the following section.

5.3.2. NAI Associated with SID

This document defines the following NAIs:

'IPv4 Node ID' is specified as an IPv4 address. In this case, ST value is 1, and the Length is 8 or 12 depending on either SID or NAI or both are included in the subobject.

'IPv6 Node ID' is specified as an IPv6 address. In this case, ST and Length are 2, and Length is 8, 20, or 24 depending on either SID or NAI or both are included in the subobject.

'IPv4 Adjacency' is specified as a pair of IPv4 addresses. In this case, ST value is 3. The Length is 8, 12, or 16 depending on either SID or NAI or both are included in the subobject, and the format of the NAI is shown in the following figure:

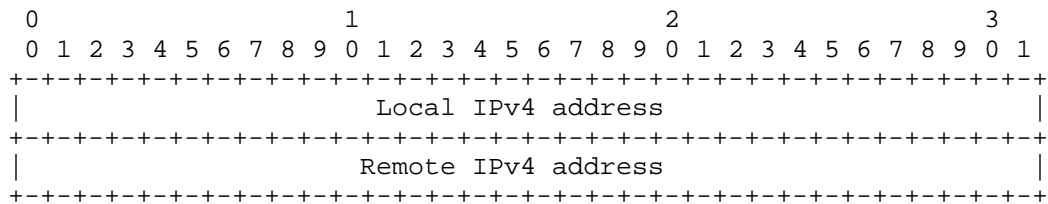


Figure 3: NAI for IPv4 Adjacency

'IPv6 Adjacency' is specified as a pair of IPv6 addresses. In this case, ST value is 4. The Length is 8, 36 or 40 depending on whether SID or NAI or both included in the subobject, and the format of the NAI is shown in the following figure:

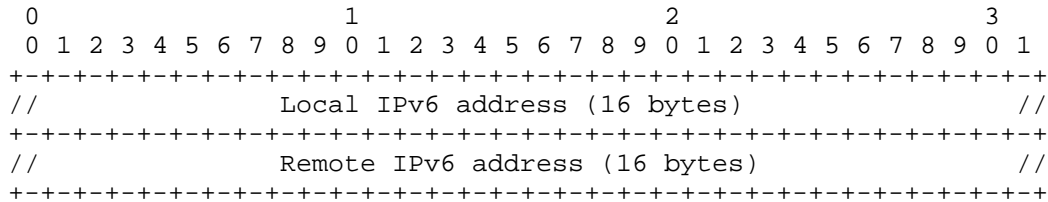


Figure 4: NAI for IPv6 adjacency

'Unnumbered Adjacency with IPv4 NodeIDs' is specified as a pair of Node ID / Interface ID tuples. In this case, ST value is 5. The Length is 8, 20, or 24 depending on whether SID or NAI or both included in the subobject, and the format of the NAI is shown in the following figure:

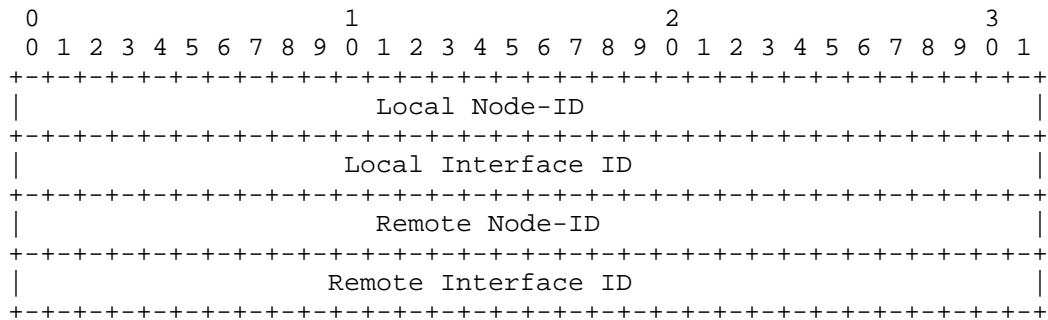


Figure 5: NAI for Unnumbered adjacency with IPv4 Node IDs

Editorial Note: We are yet to decide if another SID subobject is required for unnumbered adjacency with 128 bit node ID.

5.3.3. ERO Processing

A PCEP speaker that does not recognize the SR-ERO subobject in PCRep, PCInitiate, PCUpd or PCRpt messages MUST reject the entire PCEP message and MUST send a PCE error message with Error-Type=3 ("Unknown Object") and Error-Value=2 ("Unrecognized object Type") or Error-Type=4 ("Not supported object") and Error-Value=2 ("Not supported object Type"), defined in [RFC5440].

When the SID represents an MPLS label (i.e. the M bit is set), its value (20 most significant bits) MUST be larger than 15, unless it is special purpose label, such as an Entropy Label Indicator (ELI) or an Entropy Label (EL). If a PCEP speaker receives a label ERO subobject with an invalid value, it MUST send the PCE error message with Error-Type = 10 ("Reception of an invalid object") and Error Value = TBD ("Bad label value"). If both M and C bits of an ERO subobject are set, and if a PCEP speaker finds erroneous setting in one or more of TC, S, and TTL fields, it MUST send a PCE error with Error-Type = 10 ("Reception of an invalid object") and Error-Value = TBD ("Bad label format").

If a PCC receives a stack of SR-ERO subobjects, and the number of stack exceeds the maximum number of SIDs that the PCC can impose on the packet, it MAY send a PCE error with Error-Type = 10 ("Reception of an invalid object") and Error-Value = TBD ("Unsupported number of Segment ERO subobjects").

When a PCEP speaker detects that all subobjects of ERO are not identical, and if it cannot handle such ERO, it MUST send PCE error with Error-Type = 10 ("Reception of an invalid object") and Error-Value = TBD ("Non-identical ERO subobjects").

If a PCEP speaker receives an SR-ERO subobject in which both SID and NAI are absent, it MUST consider the entire ERO object invalid and send a PCE error with Error-Type = 10 ("Reception of an invalid object") and Error-Value = TBD ("Both SID and NAI are absent in ERO subobject").

5.4. RRO Object

A PCC can record SR-TE LSP and report the LSP to a PCE via RRO. An RRO object contains one or more subobjects called "SR-RRO subobjects" whose format is shown below:

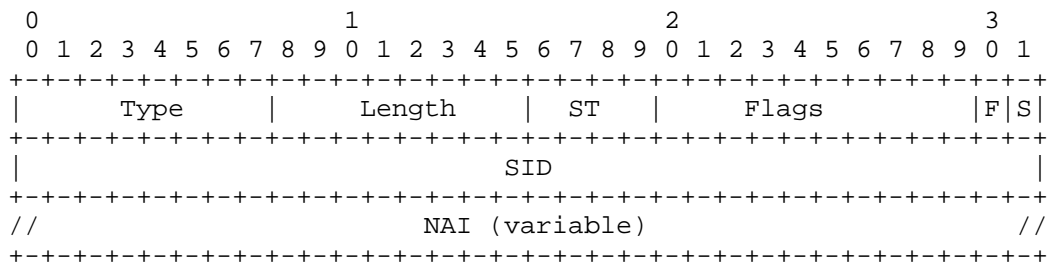


Figure 6: SR-RRO Subobject format

The format of SR-RRO subobject is the same as that of SR-ERO subobject without L, C, and M flags. The F and S flags are used with the same meaning.

A PCC MUST assume that SR-RRO subobjects are organized such that the first subobject relative to the beginning of RRO contains the information about the topmost label, and the last subobject contains information about the bottommost label of the SR-TE LSP.

5.4.1. RRO Processing

Processing rules of SR-RRO subobject are identical to those of SR-ERO subobject.

If a PCEP speaker receives an SR-RRO subobject in which both SID and NAI are absent, it MUST consider the entire RRO object invalid and send a PCE error with Error-Type = 10 ("Reception of an invalid object") and Error-Value = TBD ("Both SID and NAI are absent in RRO subobject").

6. Backward Compatibility

A PCEP speaker that does not support the SR PCEP capability cannot recognize the SR-ERO or SR-RRO subobjects. As such, it MUST send a PCEP error with Error-Type = 4 (Not supported object) and Error-Value = 2 (Not supported object Type) as per [RFC5440].

7. Management Considerations

7.1. Policy

PCEP implementation:

- o Can enable SR PCEP capability either by default or via explicit configuration.
- o May generate PCEP error due to unsupported number of SR-ERO or SR-RRO subobjects either by default or via explicit configuration.

7.2. The PCEP Data Model

A PCEP MIB module is defined in [I-D.ietf-pce-pcep-mib] needs be extended to cover additional functionality provided by [RFC5440] and [I-D.ietf-pce-pce-initiated-lsp]. Such extension will cover the new functionality specified in this document.

8. Security Considerations

The security considerations described in [RFC5440] and [I-D.ietf-pce-pce-initiated-lsp] are applicable to this specification. No additional security measure is required.

9. IANA Considerations

9.1. PCEP Objects

IANA is requested to allocate a new ERO subobject and a new RRO subobject types (recommended values = 5 and 6 respectively).

9.2. PCEP-Error Object

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

Error-Type	Meaning
10	Reception of an invalid object.

Error-value=2: Bad label value.
 Error-value=3: Unsupported number of Segment ERO subobjects.
 Error-value=4: Bad label format.
 Error-value=5: Non-identical ERO subobjects.
 Error-value=6: Both SID and NAI are absent in ERO subobject.
 Error-value=7: Both SID and NAI are absent in RRO subobject.

9.3. PCEP TLV Type Indicators

This document defines the following new PCEP TLVs:

Value	Meaning	Reference
26	SR-PCE-CAPABILITY	This document

9.4. New Path Setup Type

This document defines a new setup type for the PATH-SETUP-TYPE TLV as follows:

Value	Description	Reference
1	Traffic engineering path is setup using Segment Routing technique.	This document

10. Contributors

The following people contributed to this document:

- Lakshmi Sharma (Cisco Systems)

11. Acknowledgements

We like to thank Ina Minei, George Swallow, and Marek Zavodsky for the valuable comments.

12. References

12.1. Normative References

- [I-D.filsfils-rtgwg-segment-routing]
Filsfils, C., Previdi, S., Bashandy, A., Decraene, B., Litkowski, S., Horneffer, M., Milojevic, I., Shakir, R., Ytti, S., Henderickx, W., Tantsura, J., and E. Crabbe, "Segment Routing Architecture", draft-filsfils-rtgwg-segment-routing-01 (work in progress), October 2013.
- [I-D.ietf-isis-segment-routing-extensions]
Previdi, S., Filsfils, C., Bashandy, A., Gredler, H., Litkowski, S., and J. Tantsura, "IS-IS Extensions for Segment Routing", draft-ietf-isis-segment-routing-extensions-00 (work in progress), April 2014.
- [I-D.ietf-ospf-segment-routing-extensions]
Psenak, P., Previdi, S., Filsfils, C., Gredler, H., Shakir, R., Henderickx, W., and J. Tantsura, "OSPF Extensions for Segment Routing", draft-ietf-ospf-segment-routing-extensions-00 (work in progress), June 2014.
- [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-01 (work in progress), June 2014.
- [I-D.ietf-pce-pcep-mib]
Koushik, K., Stephan, E., Zhao, Q., King, D., and J. Hardwick, "PCE communication protocol (PCEP) Management Information Base", draft-ietf-pce-pcep-mib-04 (work in progress), February 2013.
- [I-D.ietf-pce-stateful-pce]
Crabbe, E., Medved, J., Minei, I., and R. Varga, "PCEP Extensions for Stateful PCE", draft-ietf-pce-stateful-pce-05 (work in progress), July 2013.
- [I-D.sivabalan-pce-lsp-setup-type]
Sivabalan, S., Medved, J., Minei, I., Varga, R., and E. Crabbe, "LSP setup method in PCEP messages", draft-sivabalan-pce-lsp-setup-type-00 (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.

- [RFC5462] Andersson, L. and R. Asati, "Multiprotocol Label Switching (MPLS) Label Stack Entry: "EXP" Field Renamed to "Traffic Class" Field", RFC 5462, February 2009.

12.2. Informative References

- [RFC3209] Awduche, D., Berger, L., Gan, D., Li, T., Srinivasan, V., and G. Swallow, "RSVP-TE: Extensions to RSVP for LSP Tunnels", RFC 3209, December 2001.
- [RFC3473] Berger, L., "Generalized Multi-Protocol Label Switching (GMPLS) Signaling Resource ReserVation Protocol-Traffic Engineering (RSVP-TE) Extensions", RFC 3473, January 2003.
- [RFC3477] Kompella, K. and Y. Rekhter, "Signalling Unnumbered Links in Resource ReSerVation Protocol - Traffic Engineering (RSVP-TE)", RFC 3477, January 2003.
- [RFC4657] Ash, J. and J. Le Roux, "Path Computation Element (PCE) Communication Protocol Generic Requirements", RFC 4657, September 2006.

Authors' Addresses

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

Email: msiva@cisco.com

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

Email: jmedved@cisco.com

Clarence Filsfils
Cisco Systems, Inc.
Pegasus Parc
De kleetlaan 6a, DIEGEM BRABANT 1831
BELGIUM

Email: cfilsfil@cisco.com

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

Email: edward.crabbe@gmail.com

Robert Raszuk
NTT I3
101 S. Ellsworth Ave
San Mateo, CA 94401
US

Email: robert@raszuk.net

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

Email: vlopez@tid.es

Jeff Tantsura
Ericsson
300 Holger Way
San Jose, CA 95134
USA

Email: jeff.tantsura@ericsson.com

PCE Working Group
Internet-Draft
Intended status: Standards Track
Expires: November 10, 2014

Q. Wu
D. Dhody
Huawei
D. King
Old Dog Consulting
D. Lopez
Telefonica I+D
J. Tantsura
Ericsson
May 9, 2014

Path Computation Element (PCE) Discovery using Domain Name System(DNS)
draft-wu-pce-dns-pce-discovery-06

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 discovery along with certain relevant information 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 November 10, 2014.

Copyright Notice

Copyright (c) 2014 IETF Trust and the persons identified as the document authors. All rights reserved.

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

Table of Contents

1. Introduction	3
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. PCE Virtualization	7
3.4. Additional Capabilities	7
3.4.1. Handling Changes in PCE Identities	7
3.4.2. Secure Inter-domain Discovery	8
3.4.3. Load Sharing of Path Computation Requests	8
4. Extended Naming Authority Pointer (NAPTR)Service Field Format	9
4.1. IETF Standards Track PCE Applications	10
5. Backwards Compatibility	11
6. Discovering a Path Computation Element	12
6.1. Determining the PCE Service and transport protocol	13
6.2. Determining the IP Address of the PCE	13
6.2.1. Examples	15
6.3. Determining the PCE domains and Neighbor PCE domains	16
7. IANA Considerations	17
7.1. IETF PCE Application Service Tags	17
7.2. PCE Application Protocol Tags	17
8. Security Considerations	18
9. Acknowledgements	19
10. References	20
10.1. Normative References	20
10.2. 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 same 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. (sec 4 of [PCE-QUESTION]).

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;
- o The set of communication and path computation-specific capabilities.

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 very well 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-LS: [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. PCE Virtualization

Server virtualization has gain importance since it provides better reliability and high availability in the event of hardware failure. It allows for higher utilization of physical resources while improving administration by having a single management interface for all virtual servers.

When one PCE instance is virtually hosted on a server and initiated as a PCE instance, another PCE instance may be created on the same server or a different server to provide better load balancing and reliability. In such a case, where there are a large number of PCCs that need to know these PCE instances' location, manual configuration on PCCs for PCC and PCE relationship is not trivial or desirable.

3.4. Additional Capabilities

3.4.1. Handling Changes in PCE Identities

In the case of H-PCE ,when a dynamic Address is assigned to the parent PCE, any existing configuration entry on child PCE becomes invalid and the parent PCE becomes unreachable. In order to handle changes in parent PCE identities, the DNS update can be used to provide IP reachability to the parent PCE with new assigned Address. The DNS update can be performed by either parent PCE or OSS/NMS that is aware of PCE Identities changes.

3.4.2. Secure Inter-domain Discovery

Applications make use of DNS lookups on FQDN to find a node(e.g., PCEP endpoint). When a PCE performs DNS lookup or dynamic DNS update with the DNS server, the PCE MUST have a security association of some type with the DNS server. The security association SHOULD be established either using DNSSEC [RFC4033] or TSIG/TKEY[RFC2845][RFC2930]. DNS lookup for PCE Discovery can be applied either within an administration domain or spanning across administration domains. A security association is REQUIRED even if the DNS server is in the same administrative domain as the PCE.

3.4.3. 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.

4. 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 multiple PCE applications as defined below.

```

iana-registered-service =/ pce-service
pce-service             = "pce" *("+" appln-name)
appln-name              = non-ws-string
non-ws-string           = 1*(%x21-FF)

```

The appln-name element is the Application Identifier used to identify a specific PCE application. The PCE Application Name 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]).

4.1. 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+gco:pce.tcp'`

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

5. 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.

6. 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.

6.1. 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.

6.2. 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.

6.2.1. 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+gco:pce.tcp" ""
        server1.ex2.example.com
IN NAPTR 150 50 "a" "pce+gco: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.

6.3. Determining 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 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 may follow one 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   "pce domain = as10"
ex2.example.com  IN   TXT   "neigh domain= as5"
ex2.example.com  IN   TXT   "cap=link constraint"

```

The client 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.

7. IANA Considerations

7.1. 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+gco	GCO [RFC5557]
pce+p2mp	P2MP [RFC5671]
pce+stateful	Stateful [STATEFUL-PCE]
pce+gmpls	GMPLS [RFC7025]
pce+interas	Inter-AS[RFC5376]
pce+interarea	Inter-Area [RFC4927]
pce+interlayer	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.

7.2. 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

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].

8. 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.

9. 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.

10. References

10.1. 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.

10.2. Informative References

- [ALTO] Kiesel, S., "ALTO Server Discovery", ID draft-ietf-alto-server-discovery-22, December 2013.
- [BGP-LS] Gredler, H., "North-Bound Distribution of Link-State and TE Information using BGP", ID draft-ietf-idr-ls-distribution-04, November 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

Jeff Tantsura
Ericsson
300 Holger Way
San Jose, CA 95134
US

Email: Jeff.Tantsura@ericsson.com

PCE Working Group
Internet-Draft
Intended status: Standards Track
Expires: December 26, 2014

Q. Wu
D. Dhody
Huawei
S. Previdi
Cisco Systems, Inc
June 24, 2014

Extensions to Path Computation Element Communication Protocol (PCEP) for
handling the Link Bandwidth Utilization
draft-wu-pce-pcep-link-bw-utilization-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 Link bandwidth utilization (the total bandwidth of a link in current use for the forwarding) is an important factor to consider during path computation. [OSPF-TE-EXPRESS] and [ISIS-TE-EXPRESS] define mechanisms that distribute this information via OSPF and ISIS respectively. This document describes extensions to PCEP to use 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 December 26, 2014.

Copyright Notice

Copyright (c) 2014 IETF Trust and the persons identified as the document authors. All rights reserved.

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

Table of Contents

1. Introduction	3
1.1. Requirements Language	3
2. Terminology	3
3. Link Bandwidth Utilization (LBU)	4
4. Link Reserved Bandwidth Utilization (LRBU)	4
5. PCEP Requirements	5
6. PCEP Extensions	5
6.1. BU Object	5
6.1.1. Elements of Procedure	6
6.2. New Objective Functions	7
6.3. PCEP Message Extension	8
6.3.1. The PCReq message	8
6.3.2. The PCRep message	9
7. Other Considerations	10
7.1. Reoptimization Consideration	10
7.2. Inter-domain Consideration	10
7.2.1. Inter-AS Link	10
7.3. P2MP Consideration	10
7.4. Stateful PCE	10
7.4.1. PCEP Message Extension	11
7.4.1.1. The PCRpt message	11
8. IANA Considerations	11
8.1. New PCEP Object	11
8.2. BU Object	12
8.3. Objective Functions	12
9. Security Considerations	12
10. Manageability Considerations	12
10.1. Control of Function and Policy	12
10.2. Information and Data Models	12
10.3. Liveness Detection and Monitoring	13
10.4. Verify Correct Operations	13
10.5. Requirements On Other Protocols	13
10.6. Impact On Network Operations	13
11. Acknowledgments	13
12. References	13
12.1. Normative References	13

12.2. Informative References	13
Appendix A. Contributor Addresses	15

1. Introduction

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

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

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

[OSPF-TE-EXPRESS] and [ISIS-TE-EXPRESS] also include parameters related to link latency, latency variation and packet loss. [PCE-SERVICE-AWARE] describes extensions to PCEP to consider them.

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).

LBU: Link Bandwidth Utilization. (See Section 3.)

LRBU: Link Reserved Bandwidth Utilization. (See Section 4.)

MRUP: Maximum Reserved Under-Utilized Path. (See Section 6.2.)

MUP: Maximum Under-Utilized Path. (See Section 6.2.)

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].)

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

PCEP: Path Computation Element Communication 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 (Utilized Bandwidth in [OSPF-TE-EXPRESS] and [ISIS-TE-EXPRESS]). 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 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, the Available bandwidth and LBU. The actual bandwidth by non-RSVP TE traffic can be calculated by subtracting the Available Bandwidth from the 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 $(LRBU / (\text{Maximum reservable bandwidth})) * 100$.

5. PCEP Requirements

The following requirements associated with the bandwidth utilization are identified for PCEP:

1. The PCE supporting this document MUST have the capability to compute end-to-end path with the bandwidth utilization constraints. It MUST also support the combination of the bandwidth utilization constraint with the existing constraints (cost, hop-limit...).
2. The PCC MUST be able to request for the bandwidth utilization constraint in PCReq message as the upper limit that should not be crossed for each link in the path.
3. The PCC MUST be able to request for the bandwidth utilization constraint in PCReq message as an Objective function (OF) [RFC5541] to be optimized.
4. PCEs are not required to support the 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 the bandwidth utilization constraint.
5. PCEP SHOULD provide a mechanism to handle the 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] to meet requirements outlined in Section 5. The proposed solution is used to consider the bandwidth utilization during path computation.

6.1. BU Object

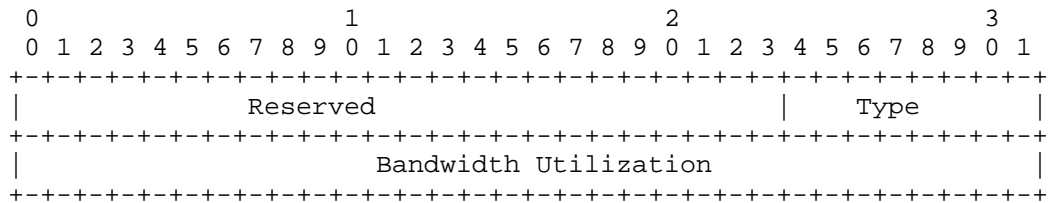
The BU (the Bandwidth Utilization) is used to indicate the upper limit of the acceptable link bandwidth utilization percentage.

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

BU Object-Class is TBD.

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. Link Bandwidth Utilization (LBU) Type is 1 and Link Reserved Bandwidth Utilization (LRBU) Type is 2.

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 8 bytes.

6.1.1. Elements of Procedure

A 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 type. If, for a given request, two or more instances of a BU object with the same 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 the 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 the 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 (the Maximum Under-Utilized Path) and MRUP (the 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 $\{L_i, (i=1\dots N)\}$.
- o A path P is a list of K links $\{L_{pi}, (i=1\dots K)\}$.
- 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 Reserved bandwidth on link L is denoted $R(L)$.

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

Objective Function Code: TBD

Name: Maximum Reserved Under-Utilized Path (MRUP)

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

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

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

6.3. PCEP Message Extension

6.3.1. The PCReq 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>]

```

6.3.2. The PCRep message

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 an LSP by monitoring changes in the bandwidth utilization parameters of one or more links on the path in the TED. In case of drastic change, it MAY ask PCE for reoptimization as per [RFC5440].

7.2. Inter-domain Consideration

[RFC5441] describes the Backward-Recursive PCE-Based Computation (BRPC) procedure to compute end to end optimized inter-domain path by cooperating PCEs. The new BU object defined in this document can be applied to end to end path computation, in similar manner as existing METRIC object.

All domains should have the same understanding of the BU object for end-to-end inter-domain path computation to make sense.

7.2.1. Inter-AS Link

The IGP in each neighbor domain can advertise its inter-domain TE link capabilities, this has been described in [RFC5316] (ISIS) and [RFC5392] (OSPF). The bandwidth related network performance link properties are described in [OSPF-TE-EXPRESS] and [ISIS-TE-EXPRESS], the same properties must be advertised using the mechanism described in [RFC5392] (OSPF) and [RFC5316] (ISIS).

7.3. P2MP Consideration

They are currently out of scope of this document.

7.4. Stateful PCE

[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 let PCCs relinquish control over some LSPs to the PCE.

The passive stateful PCE implementation MAY use the extension of PCReq and PCRep messages as defined in Section 6.3.1 and Section 6.3.2 to enable the use of BU object.

The additional objective functions defined in this document can also be used with stateful PCE.

7.4.1. PCEP Message Extension

7.4.1.1. 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 or delegate control of an LSP. The PCRpt message is extended to support BU object. This optional BU object can specify the upper limit that should not be crossed.

As per [STATEFUL-PCE], 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>
```

```
<path> ::= <ERO><attribute-list>[<RRO>]
```

Where <attribute-list> is extended as per Section 6.3.2 for BU object.

Thus a BU object can be used to specify the upper limit set at the PCC at the time of LSP delegation to an active stateful PCE.

8. IANA Considerations

IANA assigns values to PCEP parameters in registries defined in [RFC5440]. IANA has made the following additional assignments.

8.1. New PCEP Object

IANA assigned a new object class in the registry of PCEP Objects as follows.

Object Class	Object Type	Name	Reference

TBD	1	BU	[This I.D.]

8.2. BU Object

IANA created a registry to manage the codespace of the Type field of the METRIC Object.

Codespace of the T field (Metric Object)

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

8.3. Objective Functions

Two new Objective Functions have been defined. IANA has made the following allocations from the PCEP "Objective Function" sub-registry:

Code Point	Name	Reference
TBA	Maximum Under-Utilized Path (MUP)	[This I.D.]
TBA	Maximum Reserved Under-Utilized Path (MRUP)	[This I.D.]

9. Security Considerations

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

10. Manageability Considerations

10.1. 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.

10.2. Information and Data Models

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

10.3. 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].

10.4. Verify Correct Operations

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

10.5. 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].

10.6. 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].

11. Acknowledgments

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

12. References

12.1. Normative References

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

12.2. Informative References

[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.

[RFC5316] Chen, M., Zhang, R., and X. Duan, "ISIS Extensions in Support of Inter-Autonomous System (AS) MPLS and GMPLS Traffic Engineering", RFC 5316, December 2008.

- [RFC5392] Chen, M., Zhang, R., and X. Duan, "OSPF Extensions in Support of Inter-Autonomous System (AS) MPLS and GMPLS Traffic Engineering", RFC 5392, January 2009.
- [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.
- [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-05 (work in progress), December 2013.
- [ISIS-TE-EXPRESS]
Previdi, S., Giacalone, S., Ward, D., Drake, J., Atlas, A., Filsfils, C., and W. Wu, "IS-IS Traffic Engineering (TE) Metric Extensions", draft-ietf-isis-te-metric-extensions-03 (work in progress), April 2014.
- [PCEP-MIB]
Koushik, K., Emile, S., Zhao, Q., King, D., and J. Hardwick, "Path Computation Element Protocol (PCEP) Management Information Base", draft-ietf-pce-pcep-mib-08 (work in progress), April 2014.
- [STATEFUL-PCE]
Crabbe, E., Medved, J., Minei, I., and R. Varga, "PCEP Extensions for Stateful PCE", draft-ietf-pce-stateful-pce-08 (work in progress), February 2014.
- [PCE-SERVICE-AWARE]
Dhody, D., Manral, V., Ali, Z., Swallow, G., and K. Kumaki, "Extensions to the Path Computation Element Communication Protocol (PCEP) to compute service aware Label Switched Path (LSP).", draft-ietf-pce-pcep-service-aware-04 (work in progress), March 2014.

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

Zafar Ali
Cisco Systems

EMail: zali@cisco.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

PCE Working Group
Internet-Draft
Intended status: Standards Track
Expires: December 28, 2014

Q. Wu
D. Dhody
Huawei
M. Boucadair
C. Jacquenet
France Telecom
J. Tantsura
Ericsson
June 26, 2014

PCEP Extensions for traffic steering support in Service Function
Chaining
draft-wu-pce-traffic-steering-sfc-04

Abstract

This document provides an overview of the usage of Path Computation Element (PCE) with Service Function Chaining (SFC); which is described as the definition and instantiation of an ordered set of such service functions (such as firewalls, load balancers), and the subsequent "steering" of traffic flows through those service functions.

Further this document specifies extensions to the Path Computation Element Protocol (PCEP) that allow a stateful PCE to compute and instantiate Service Function Paths (SFP).

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 28, 2014.

Copyright Notice

Copyright (c) 2014 IETF Trust and the persons identified as the document authors. All rights reserved.

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

Table of Contents

1. Introduction	2
2. Conventions used in this document	3
3. Service Function Paths and PCE	3
4. Overview of PCEP Operation in SFC enabled Networks	5
4.1. SFP Instantiation	5
4.2. SFP Deletion	5
4.3. SFP Delegation and Cleanup	5
4.4. SFP State Synchronization	5
4.5. SFP Update and Report	5
5. Object Formats	6
5.1. The OPEN Object	6
5.2. The LSP Object	6
5.2.1. SFP Identifiers TLV	7
6. Backward Compatibility	7
7. Relationship to SR	7
8. Security Considerations	7
9. IANA Considerations	8
10. References	8
10.1. Normative References	8
10.2. Informative References	8

1. Introduction

Service chaining enables creation of composite services that consist of an ordered set of Service Functions (SF) that must be applied to packets and/or frames selected as a result of classification as described in [I-D.boucadair-sfc-framework][I-D.quinn-sfc-arch] and referred to as Service Function Chain (SFC). Service Function Path (SFP) is the instantiation of a SFC in the network. Packets follow a Service Function Path from a classifier through the requisite Service Functions (SF).

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

[I-D.ietf-pce-stateful-pce] specifies extensions to PCEP to enable stateful control of MPLS TE LSPs. [I-D.ietf-pce-pce-initiated-lsp] provides the fundamental extensions needed for stateful PCE-initiated LSP instantiation.

This document specifies extensions to the PCEP that allow a stateful PCE to compute and instantiate Service Function Paths (SFP).

2. Conventions used in this document

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

The following terminologies are used in this document:

PCC: Path Computation Client.

PCE: Path Computation Element.

PCEP: Path Computation Element Protocol.

PDP: Policy Decision Point.

SF: Service Function.

SFC: Service Function Chain.

SFP: Service Function Path.

UNI: User-Network Interface.

3. Service Function Paths and PCE

Services are constructed as a sequence of SFs that represent an SFC, where SF can be a virtual instance or be embedded in a physical network element, and one or more SFs may be deployed within the same physical network element. SFC creates an abstracted view of a service and specifies the set of required SFs as well as the order in which they must be executed.

When an SFC is instantiated into the network it is necessary to select the specific instances of SFs that will be used, and to create the service topology for that SFC using SF's network locator. Thus, instantiation of the SFC results in the creation of a Service Function Path (SFP) and is used for forwarding packets through the SFC. In other words, an SFP is the instantiation of the defined SFC as described in details in [I-D.boucadair-sfc-framework][I-D.quinn-sfc-arch].

The selection of SFP can be based on a range of policy attributes, ranging from simple to more elaborate criteria and stateful PCE with extensions to PCEP are one such way to achieve this.

Stateful pce [I-D.ietf-pce-stateful-pce] specifies a set of extensions to PCEP to enable stateful control of TE LSPs. [I-D.ietf-pce-pce-initiated-lsp] provides the fundamental motivations and extensions needed for stateful PCE-initiated LSP instantiation. This document specifies extensions that allow a stateful PCE to compute and instantiate Service Function Paths (SFP) via PCEP.

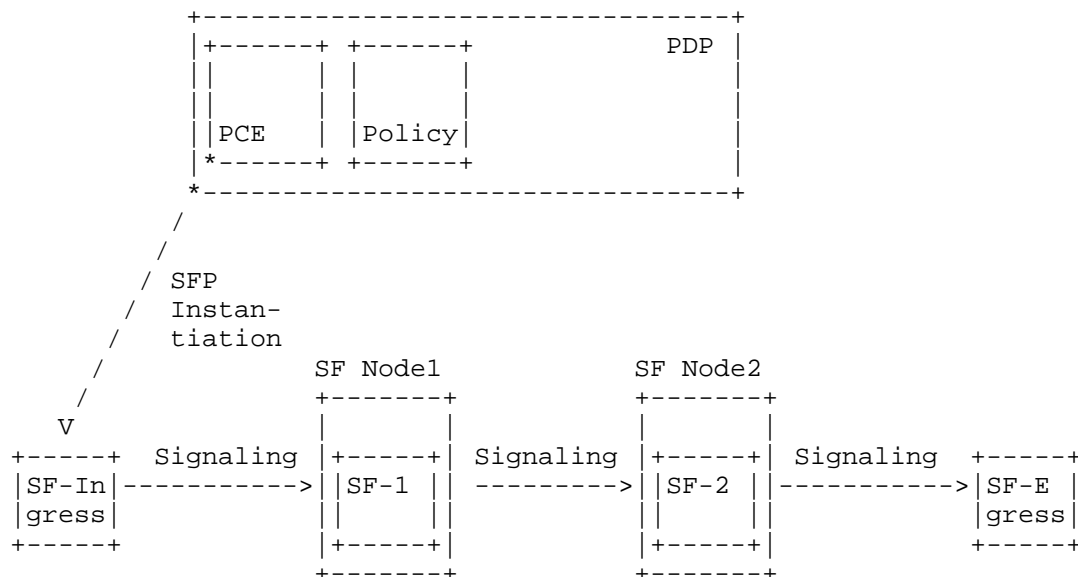


Figure 1: SFP instantiation vis PCE

A Policy Decision Point (PDP) [RFC2753] is the central entity which is responsible for maintaining SFC Policy Tables and enforcing appropriate policies in SF Nodes described in detail in [I-D.boucadair-sfc-framework]. A PDP may further use stateful PCE and its instantiation mechanism to compute and instantiate Service

Function Paths (SFP). The PCE maybe co-located with the PDP or an external entity.

4. Overview of PCEP Operation in SFC enabled Networks

A PCEP speaker indicates its ability to support PCE initiated dynamic SFP during the PCEP Initialization Phase via mechanism described in Section 5.1.

As per section 5.1 of [I-D.ietf-pce-pce-initiated-lsp], the PCE sends a Path Computation LSP Initiate Request (PCInitiate) message to the PCC to instantiate or delete a LSP. This document makes no change to the PCInitiate message format but extends LSP objects described in Section 5.2.

4.1. SFP Instantiation

The Instantiation operation of SFP is same as defined in section 5.3 of [I-D.ietf-pce-pce-initiated-lsp]. Rules of processing and error codes remains unchanged.

4.2. SFP Deletion

The deletion operation of SFP is same as defined in section 5.4 of [I-D.ietf-pce-pce-initiated-lsp] by sending an LSP Initiate Message with an LSP object carrying the PLSP-ID of the SFP to be removed and an SRP object with the R flag set (LSP-REMOVE as per section 5.2 of [I-D.ietf-pce-pce-initiated-lsp]). Rules of processing and error codes remains unchanged.

4.3. SFP Delegation and Cleanup

SFP delegation and cleanup operations are same as defined in section 6 of [I-D.ietf-pce-pce-initiated-lsp]. Rules of processing and error codes remains unchanged.

4.4. SFP State Synchronization

State Synchronization operations described in Section 5.4 of [I-D.ietf-pce-stateful-pce] and can be applied for SFPs as well.

4.5. SFP Update and Report

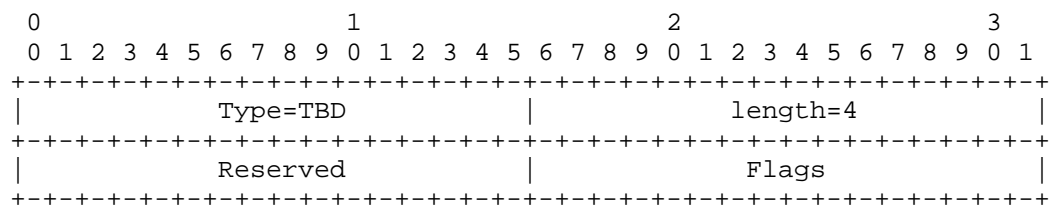
PCE can make an SFP Update requests to a PCC to update one or more attributes of an SFP and to re-signal the SFP with updated attributes. PCC can make an SFP state report to a PCE to send SFP state. The mechanism are described in [I-D.ietf-pce-stateful-pce] and can be applied for SFPs as well.

5. Object Formats

5.1. The OPEN Object

This document defines a new optional TLV for use in the OPEN Object to indicate the PCEP speaker's capability for Service function Chaining.

The SFC-PCE-CAPABILITY TLV is an optional TLV for use in the OPEN Object to advertise the SFC capability on the PCEP session. The format of the SFC-PCE-CAPABILITY TLV is shown in the following figure:



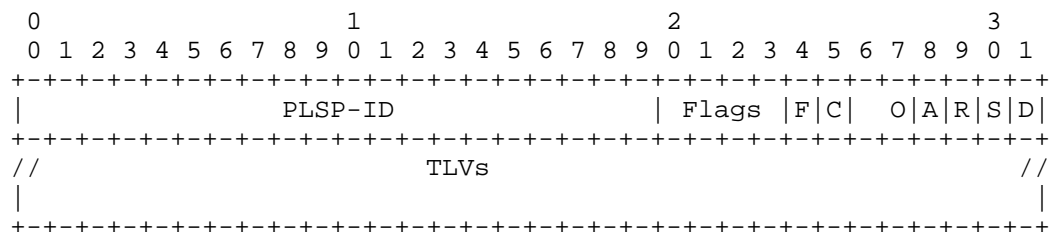
The code point for the TLV type is to be defined by IANA. The TLV length is 4 octets.

The value is TBD.

As per [I-D.ietf-pce-stateful-pce], PCEP speaker advertises capability for instantiation of PCE-initiated LSPs via Stateful PCE Capability TLV (LSP-INSTANTIATION-CAPABILITY bit) in open message. The inclusion of SFC-PCE-CAPABILITY TLV in an OPEN object indicates that the sender is SFC capable. These mechanism when used together indicates the instantiation capability for SFP by the PCEP speaker.

5.2. The LSP Object

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



A new flag, the SFC (F) flag is introduced. The F Flag set to 1 to indicate that this an SFP. The C flag will also be set to indicate it was created via a PCInitiate message.

5.2.1. SFP Identifiers TLV

The SFP Identifiers TLV MUST be included in the LSP object for Service Function Paths (SFP).

The format and operations are TBD.

6. Backward Compatibility

The PCEP protocol extensions described in this document for PCEP speaker with instantiation capability for SFPs MUST NOT be used if PCC or PCE has not advertised its stateful capability with Instantiation and SFC capability as per Section 5.1. If this is not the case and Stateful operations on SFPs are attempted, then a PCErr with error-type 19 (Invalid Operation) and error-value TBD needs to be generated.

[Editor Note: more information on exact error value is needed]

7. Relationship to SR

Segment Routing (SR) technology leverages the source routing and tunneling paradigms where a source node can choose a path without relying on hop-by-hop signaling. A stateful PCE can be used 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.sivabalan-pce-segment-routing].

The SFP instantiation mechanism described in this document is not tightly coupled to any SFP signaling mechanism. Thus SR based SFP can also utilize the mechanism described here and do not need another set of protocol extensions.

8. Security Considerations

The security considerations described in [RFC5440] and [I-D.ietf-pce-pce-initiated-lsp] are applicable to this specification. No additional security measure is required.

9. IANA Considerations

TBD

10. References

10.1. Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [I-D.ietf-pce-stateful-pce]
Crabbe, E., Minei, I., Medved, J., and R. Varga, "PCEP Extensions for Stateful PCE", draft-ietf-pce-stateful-pce-09 (work in progress), June 2014.
- [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-01 (work in progress), June 2014.

10.2. Informative References

- [RFC2753] Yavatkar, R., Pendarakis, D., and R. Guerin, "A Framework for Policy-based Admission Control", RFC 2753, January 2000.
- [RFC5440] Vasseur, JP. and JL. Le Roux, "Path Computation Element (PCE) Communication Protocol (PCEP)", RFC 5440, March 2009.
- [I-D.quinn-sfc-arch]
Quinn, P. and J. Halpern, "Service Function Chaining (SFC) Architecture", draft-quinn-sfc-arch-05 (work in progress), May 2014.
- [I-D.boucadair-sfc-framework]
Boucadair, M., Jacquenet, C., Parker, R., Lopez, D., Guichard, J., and C. Pignataro, "Service Function Chaining: Framework & Architecture", draft-boucadair-sfc-framework-02 (work in progress), February 2014.
- [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.

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.ietf@gmail.com

Mohamed Boucadair
France Telecom
Rennes 35000
France

EMail: mohamed.boucadair@orange.com

Christian Jacquenet
France Telecom
Rennes 35000
France

EMail: christian.jacquenet@orange.com

Jeff Tantsura
Ericsson
300 Holger Way
San Jose, CA 95134
US

EMail: Jeff.Tantsura@ericsson.com

Pce Working Group
Internet-Draft
Intended status: Standards Track
Expires: December 24, 2014

X. Xu
J. You
Huawei
H. Shah
Ciena
L. Contreras
Telefonica I+D
June 22, 2014

PCEP Extensions for SFC in SR Networks
draft-xu-pce-sr-sfc-01

Abstract

[I-D.xu-spring-pce-based-sfc-arch] describes a PCE-based SFC architecture in which the PCE is used to compute service function paths in SR networks. Based on the above architecture, this document describes extensions to the Path Computation Element Protocol (PCEP) that allow a PCE to compute and instantiate service function paths in SR networks.

Requirements Language

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

Status of This Memo

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

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

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

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

Copyright Notice

Copyright (c) 2014 IETF Trust and the persons identified as the document authors. All rights reserved.

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

Table of Contents

1. Introduction	3
2. Terminology	4
3. Overview of PCEP Extensions for SFC in SR Networks	4
4. PCEP Message Extensions for SR-based SFC	5
4.1. PCReq Message	5
4.2. PCRep Message	5
5. Object Formats	5
5.1. OPEN Object	5
5.1.1. SR-SFC PCE Capability TLV	6
5.2. RP Object	6
5.3. Include Route Object	7
5.4. SR-SFC-ERO Object	7
5.4.1. SR-SFC-ERO Subobject	7
5.4.2. NSI Associated with SID	9
5.4.3. SR-SFC-ERO Processing	9
6. IANA Considerations	9
6.1. PCEP Objects	9
6.2. PCEP-Error Object	9
6.3. PCEP TLV Type Indicators	10
6.4. New Path Setup Type	10
6.5. New IRO Sub-object Type	10
7. Security considerations	10
8. Acknowledgement	10
9. References	10
9.1. Normative References	10
9.2. Informative References	11
Authors' Addresses	11

1. Introduction

Service Function Chaining (SFC) provides a flexible way to construct services. When applying a particular service function chain to the traffic classified by the SFC classifier, the traffic needs to be steered through an ordered set of service functions in the network. This ordered set of service functions in the network, referred to as a Service Function Path (SFP), is an instantiation of the service function chain in the network. For example, as shown in Figure 1, an SFP corresponding to the SFC of {SF1, SF3} can be expressed as {Service Node 1, SF1, Service Node 2, SF3}.

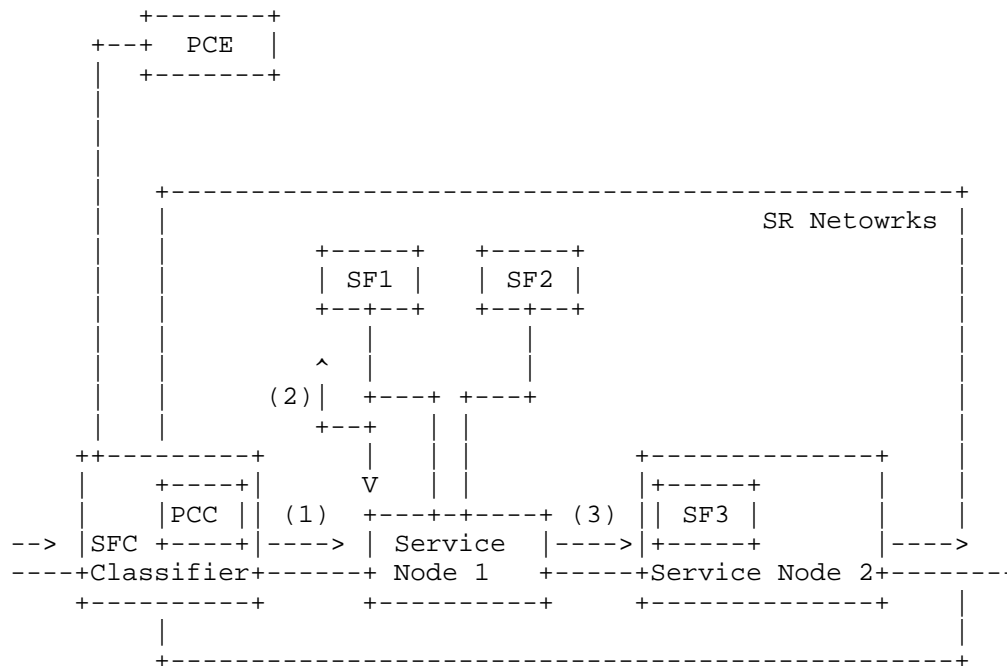


Figure 1: PCE-based Service Function Chaining in SR Network

[I-D.xu-spring-pce-based-sfc-arch] describes a PCE-based SFC architecture in which the PCE is used to compute a service function path (i.e., instantiate a service function chain) in SR networks. This document describes extensions to the PCEP based on that architecture.

2. Terminology

This section contains definitions for terms used frequently throughout this document. However, many additional definitions can be found in [RFC5440], [I-D.sivabalan-pce-segment-routing] and [I-D.xu-spring-pce-based-sfc-arch].

PCC: Path Computation Client

PCE: Path Computation Element

PCEP: Path Computation Element Protocol

ERO: Explicit Route Object

SF Identifier (SF ID): A unique identifier that represents a service function within an SFC-enabled domain.

Service Function Path (SFP): The instantiation of an SFC in the network. Specifically, it is an ordered list of service node locators and SF IDs.

Compact SFP: An ordered list of service node locators.

SID: Segment Identifier

Service Function SID : A locally unique SID indicating a particular service function on a service node.

SR: Segment Routing

SR-specific SFP: An ordered list of node SIDs (representing service nodes) and Service Function SIDs.

Compact SR-specific SFP: An ordered list of node SIDs (representing service nodes).

3. Overview of PCEP Extensions for SFC in SR Networks

As discussed in [I-D.xu-spring-pce-based-sfc-arch], the PCC provides an ordered list of SF IDs to the PCE and indicates to the PCE that what type of path is requested (e.g., an SFP, or a compact SFP, or an SR-specific SFP, or a compact SR-specific SFP), and then the PCE responds with a corresponding path.

4. PCEP Message Extensions for SR-based SFC

4.1. PCReq Message

This document does not specify any changes to the PCReq message format. This document requires the PATH-SETUP-TYPE TLV [I-D.sivabalan-pce-lsp-setup-type] to be carried in the RP Object in order for a PCC to request a particular type of path. Four new Path Setup Types need to be defined for SR-based SFC, or SR-SFC in short (Section 5.2). This document also requires the Include Route Object (IRO) to be carried in the PCReq message in order for a PCC to specify that the computed SFP must traverse a set of specified service functions. A new IRO sub-object type needs to be defined for SFC (Section 5.3).

4.2. PCRep Message

This document defines the format of the PCRep message carrying an SFP. The message is sent by a PCE to a PCC in response to a previously received PCReq message, where the PCC requested an SFP. The format of the SFC-specific PCRep message is as follows:

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

Where:

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

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

Where:

```
<path-list> ::= <SR-SFC-ERO> [<path-list>]
```

The RP and NO-PATH Objects are defined in [RFC5440]. The <SR-SFC-ERO> object contains the SFP and is defined in Section 5.4.

5. Object Formats

5.1. OPEN Object

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

5.1.1.1. SR-SFC PCE Capability TLV

The SR-SFC-PCE-CAPABILITY TLV is an optional TLV for use in the OPEN Object to negotiate SR-SFC capability on the PCEP session. The format of the SR-SFC-PCE-CAPABILITY TLV is shown in the following Figure 2:

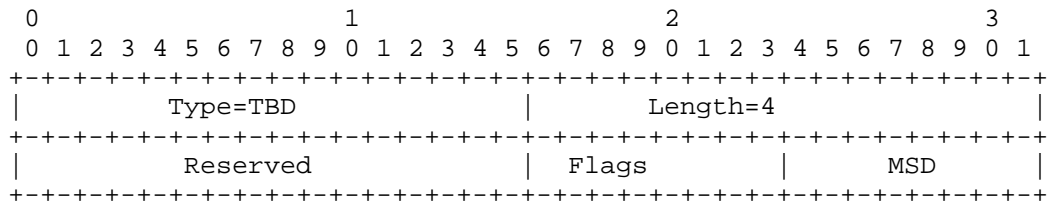


Figure 2: SR-SFC-PCE-CAPABILITY TLV format

The code point for the TLV type is to be defined by IANA. The TLV length is 4 octets. The 32-bit value is formatted as follows. The "Maximum SID Depth" (1 octet) field (MSD) specifies the maximum number of SIDs that a PCC is capable of imposing on a packet. The "Flags" (1 octet) and "Reserved" (2 octets) fields are currently unused, and MUST be set to zero and ignored on receipt.

5.1.1.1.1. Negotiating SR-SFC Capability

The SR-SFC capability TLV is contained in the OPEN object. By including the TLV in the OPEN message to a PCE, a PCC indicates its support for SFPs. By including the TLV in the OPEN message to a PCC, a PCE indicates that it is capable of computing SFPs.

5.2. RP Object

In order to setup an SFP, the RP object MUST carry a PATH-SETUP-TYPE TLV specified in [I-D.sivabalan-pce-lsp-setup-type]. This document defines four new Path Setup Types (PST) for SR-SFC as follows:

PST = 2: The path is an SFP.

PST = 3: The path is a compact SFP.

PST = 4: The path is an SR-specific SFP.

PST = 5: The path is a compact SR-specific SFP.

5.3. Include Route Object

The IRO (Include Route Object) MUST be carried within PCReq messages to indicate a particular SFC. Furthermore, the IRO MAY be carried in PCRep messages. When carried within a PCRep message with the NO-PATH object, the IRO indicates the set of service functions that cause the PCE to fail to find a path.

This document defines a new sub-object type for the SR-SFC as follows:

Type	Sub-object
5	Service Function ID

5.4. SR-SFC-ERO Object

Generally speaking, an SR-SFC-ERO object consists of one or more ERO subobjects described in the following sub-sections to represent a particular type of service function path. In the ERO subobject, each SID is associated with an identifier that represents either a service node or a service function. This identifier is referred to as the 'Node or Service Identifier' (NSI). As described later, an NSI can be represented in various formats (e.g., IPv4 address, IPv6 address, SF identifier, etc). Specifically, in the SFP case, the NSI of every ERO subobject contained in the SR-SFC-ERO object represents a service node or a service function while the SID of each ERO subobject is set to null. In the compact SFP case, the NSI of every ERO subobject contained in the SR-SFC-ERO object only represents a service node meanwhile the SID of every ERO subobject is set to null. In the SR-specific SFP, the NSI of every ERO subobject contained in the SR-SFC-ERO object represents a service node or a service function while the SID of every ERO subject MUST NOT be null. In the compact SR-specific SFP, the NSI of every ERO subobject contained in the SR-SFC-ERO object represents a service node meanwhile the SID of every ERO subobject MUST NOT be null.

5.4.1. SR-SFC-ERO Subobject

An SR-SFC-ERO subobject (as shown in Figure 3) consists of a 32-bit header followed by the SID and the NSI associated with the SID. The SID is a 32-bit or 128 bit number. The size of the NSI depends on its respective type, as described in the following sub-sections.

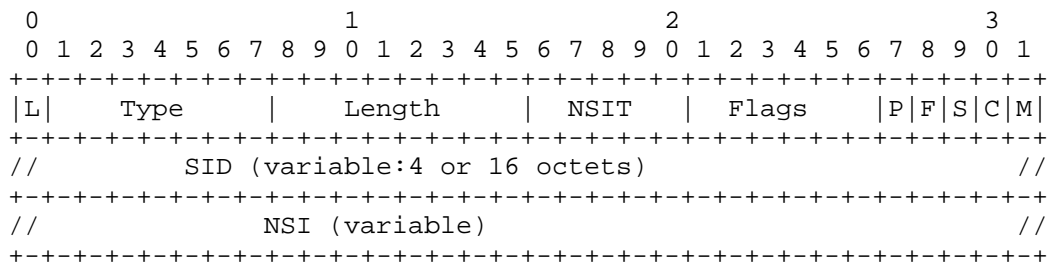


Figure 3: SR-SFC-ERO Subobject Format

The fields in the ERO Subobject are as follows:

'L' Flag: indicates whether the subobject represents a loose-hop in the explicit route [RFC3209]. If this flag is unset, a PCC MUST not overwrite the SID value present in the SR-SFC-ERO subobject. Otherwise, a PCC MAY expand or replace one or more SID value(s) in the received SR-SFC-ERO based on its local policy.

Type: is the type of the SR-SFC-ERO Subobject. This document defines the SR-SFC-ERO Subobject type. A new code point will be requested for the SR-SFC-ERO Subobject from IANA.

Length: contains the total length of the subobject in octets, including the L, Type and Length fields. Length MUST be at least 4, and MUST be a multiple of 4.

NSI Type (NSIT): indicates the type of NSI associated with the SID. The NSI-Type values are described later in this document.

Flags: is used to carry any additional information pertaining to SID. Currently, the following flag bits are defined:

M: When this bit is set, the SID value represents an MPLS label stack entry as specified in [RFC5462], where only the label value is specified by the PCE. Other fields (TC, S, and TTL) fields MUST be considered invalid, and PCC MUST set these fields according to its local policy and MPLS forwarding rules.

C: When this bit as well as the M bit are set, then the SID value represents an MPLS label stack entry as specified in [RFC5462], where all the entry's fields (Label, TC, S, and TTL) are specified by the PCE. However, a PCC MAY choose to override TC, S, and TTL values according its local policy and MPLS forwarding rules.

S: When this bit is set, the SID value in the subobject body is null. In this case, the PCC is responsible for choosing the SID value, e.g., by looking up its Traffic Engineering Database (TED) using node/service identifier in the subobject body.

F: When this bit is set, the NSI value in the subobject body is null.

P: When this bit is set, the SID value represents an IPv6 address.

SID: is the 4-octect or 16-octect Segment Identifier

NSI: contains the NSI associated with the SID. Depending on the value of NSIT, the NSI can have different format as described in the following sub-section.

5.4.2. NSI Associated with SID

This document defines the following NSIs:

'IPv4 Node ID': is specified as an IPv4 address. In this case, NSIT and Length are 1 and 12 respectively.

'IPv6 Node ID': is specified as an IPv6 address. In this case, NSIT and Length are 2 and 24 respectively.

'Service ID': is specified as an SF ID. In this case, NSIT and Length are TBD.

5.4.3. SR-SFC-ERO Processing

TBD.

6. IANA Considerations

6.1. PCEP Objects

IANA is requested to allocate an ERO subobject type (recommended value= 6) for the SR-SFC-ERO subobject.

6.2. PCEP-Error Object

TBD.

6.3. PCEP TLV Type Indicators

This document defines the following new PCEP TLVs:

Value	Meaning	Reference
27	SR-SFC-PCE-CAPABILITY	This document

6.4. New Path Setup Type

This document defines a new setup type for the PATH-SETUP-TYPE TLV as follows:

Value	Description	Reference
2	The path is an SFP.	This document
3	The path is a compact SFP.	This document
4	The path is an SR-specific SFP.	This document
5	The path is a compact SR-specific SFP.	This document

6.5. New IRO Sub-object Type

This document defines a new IRO sub-object type for the SFC as follows:

Type	Sub-object
5	Service Function ID

7. Security considerations

This document does not introduce any new security considerations.

8. Acknowledgement

TBD.

9. References

9.1. Normative References

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

- [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.
- [RFC5462] Andersson, L. and R. Asati, "Multiprotocol Label Switching (MPLS) Label Stack Entry: "EXP" Field Renamed to "Traffic Class" Field", RFC 5462, February 2009.

9.2. Informative References

- [I-D.sivabalan-pce-lsp-setup-type]
Sivabalan, S., Medved, J., Minei, I., Varga, R., and E. Crabbe, "Conveying path setup type in PCEP messages", draft-sivabalan-pce-lsp-setup-type-01 (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.xu-spring-pce-based-sfc-arch]
Xu, X., "PCE-based SFC Architecture in SR Networks", draft-xu-spring-pce-based-sfc-arch-00 (work in progress), April 2014.

Authors' Addresses

Xiaohu Xu
Huawei

Email: xuxiaohu@huawei.com

Jianjie You
Huawei
101 Software Avenue, Yuhuatai District
Nanjing, 210012
China

Email: youjianjie@huawei.com

Himanshu Shah
Ciena

Email: hshah@ciena.com

Luis M. Contreras
Telefonica I+D
Ronda de la Comunicacion, s/n
Sur-3 building, 3rd floor
Madrid, 28050
Spain

Email: luismiguel.contrerasmurillo@telefonica.com
URI: <http://people.tid.es/LuisM.Contreras/>

PCE Working Group
Internet Draft
Category: Standards track

Xian Zhang
Haomian Zheng
Huawei
Oscar Gonzales de Dios
Victor Lopez
Telefonica I+D

Expires: January 3, 2015

July 3, 2014

Extensions to Path Computation Element Protocol (PCEP) to Support
Resource Sharing-based Path Computation

draft-zhang-pce-resource-sharing-01.txt

Status of this Memo

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

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

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

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

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

This Internet-Draft will expire on January 4, 2015.

Copyright Notice

Copyright (c) 2014 IETF Trust and the persons identified as the document authors. All rights reserved.

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

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].

Abstract

Resource sharing in a network means two or more Label Switched Paths (LSPs) use common piece(s) of resource along their paths. This can help save network resource and useful in scenarios such as LSP recovery or when two LSPs do not need to be active at the same time. A Path Computation Element (PCE) is a centralized entity, responsible for path computation. Given this feature and its access to the network resource information and possibly active LSPs information, it can be used to support resource-sharing-based path computation with better efficiency.

This document extends the Path Computation Element Protocol (PCEP) in order to support resource sharing-based path computation.

Table of Contents

1. Introduction and Motivation	3
2. Motivation	4
2.1. Use Case 1	4
2.2. Use Case 2	6
3. Extensions to PCEP	7
3.1. Resource Sharing Object	8
3.2. Processing Rules	9
3.3. Carrying RSO in a PCEP Message	9
4. Security Considerations	10
5. IANA Considerations	11
5.1. New Object Type	11
6. References	12
6.1. Normative References	12

6.2. Informative References	12
7. Authors' Addresses	12

1. Introduction and Motivation

A Path Computation Element (PCE) provides an alternative way for providing path computation function, and it is especially useful in the scenarios where complex constraints and/or a demanding amount of computation resource are required [RFC4655]. The development of PCE standardization has evolved from stateless to stateful. A stateful PCE has access to the LSP database information of the network(s) it serves as a computation engine [Stateful-PCE]. Unless specified, this document assumes a PCE mentioned is a stateful PCE (either passive or active).

Resource sharing denotes that two or more Label Switched Paths (LSPs) share common piece(s) of resource, (such as a common time slot of a link in an Optical Transport Network (OTN)). This is usually useful in the scenario where only one LSP is active and the benefit herein is to save network resources. A simple example of this is dynamically calculating a LSP for an existing LSP undergoing a link failure. Note that the resource sharing can be worked out using a stateless PCE, but the mechanism may be complex and is out the scope of this draft.

This document considers the following requirement: resource sharing with one or multiple existing LSPs.

In a single domain, this is a common requirement in the recovery cases especially in order to increase traffic resilience against failure while reducing the amount of network resource used for recovery purpose [RFC4428].

The current protocol supporting the communication between a PCE and a Path Computation Client (PCC), i.e. PCE Protocol (PCEP), allows for re-optimization of an existing LSP [RFC5440]. This is achieved by setting R bit in the Request Parameter (RP) object, together with some additional information if applicable, in the Path Computation Request (PCReq) message sent from a PCC to the PCE. To support this type of resource sharing, a PCC needs to ask a PCE to compute a new path with the constraints of sharing resource with one or multiple existing LSPs. It is worth noting the 'resource sharing' in this draft not only means one LSP re-using the same link(s) of another LSP, but also the same slice of bandwidth. This may occur when an LSP is required for re-routing, or online re-optimization. Current PCEP specifications do not provide such function.

As mentioned in [stateful-PCE], the standardization of stateful PCEs also facilitates PCEP to meet this requirement since a LSP can be identified using a unique number. This simplifies configuration of PCCs by making it simpler for a PCC to request resource sharing without having to determine all of the resources to be shared.

The resource sharing can also be required across layers. This is similar to the previous requirement. However, it is more complex and therefore deserves a more detailed explanation here.

In a multi-layer network, Label Switched Paths (LSPs) in a lower layer are used to carry higher-layer LSPs across the lower-layer network [RFC5623]. Therefore, the resource sharing constraints in the higher layer might actually relate to the resource sharing in the lower layer. Thus, it is useful to consider how this can be achieved and whether additional extensions are needed using the models defined in [RFC5623].

In the next sections, use cases are provided to show what information needs to be exchanged to fulfill these requirements. This memo then provides extensions to PCEP to enable this function.

2. Motivation

2.1. Use Case 1

Figure 1 shows a single domain network with a stateful PCE. Assume a working LSP (N1-N2-N3) exists in the network. When there is failure on the link N2-N3, it is desired to set up a restoration path for this working LSP. Suppose N1 serves as the PCC and sends a request to the stateful PCE for such an LSP. Before sending the request, N1 may need to check what policy is configured locally on N1. For example, it might value resource sharing and prefer to share as much resource with the working LSP as possible and specify this in the PCReq message. Effectiveness here denotes whether the traffic can be diverted back to the working LSP immediately once the failure on the working LSP is repaired. In the case where resource sharing is more important, it would prefer to share as much resource with the working LSP as possible and specify this in the PCReq message.

On the other hand, in some case the LSP should be restored without any interruption with best effort, for example the online re-optimization. In such cases, it would prefer to share as few resources as possible. The best result for such case is to find a separate path that can make the LSP before break, which means no resource sharing involved. This can actually be implemented with

existing PCEP mechanism. However, if there is no such separate path, existing PCEP will reply error. A secondary option for this case is to set up an LSP and complete such re-optimization with resource sharing, even if some interruption introduced. Given the resource from the LSP to be interrupted, there may be some solutions instead of Path Compute error due to the lack of resource.

A simple illustration is provided below:

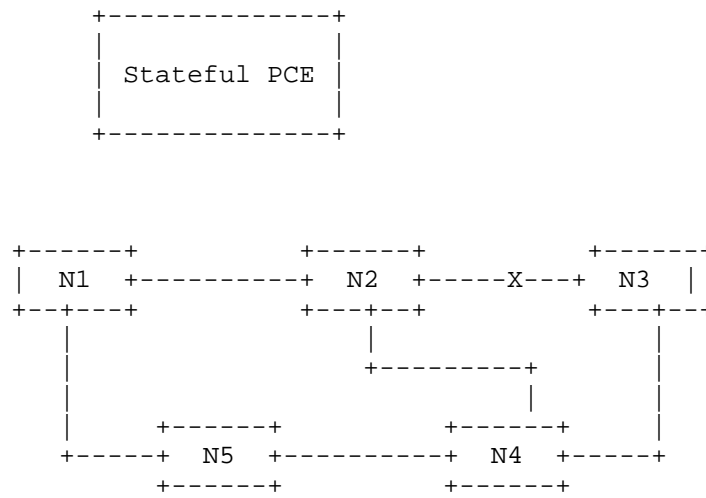


Figure 1: A Single Domain Example

Available recovery paths computed by the stateful PCE:

LSP1: N1-N2-N4-N3

LSP2: N1-N5-N4-N3

If resource sharing is preferred, the stateful PCE will reply with LSP1 information. Instead, if effectiveness is valued higher, it will reply with LSP2 information.

Another piece of information that needs to be conveyed to the PCE is the information about the working path LSP. Note this simple use case assumes end-to-end recovery. But in order to be applicable to use cases such as shared mesh protection purpose, where the head-end or tail-end nodes may be different, this information is necessary in the message exchange between PCCs and PCEs, so that the stateful PCE knows which LSP the path computation request wants to share the resource.

Besides, parameter changes during the resource sharing computation also need to be considered. For example, the bandwidth of the request may be different with the existing LSP, but still ask for resource sharing. PCE should consider the sharing request together with the policy and available resource(s) in the network. Details can be found in Section 3.3.

2.2. Use Case 2

Figure 2 shows a two-layer network example, with each layer managed by a PCE (shown in the graph as PCE Hi (for higher layer) and PCE Lo for PCE lower layer). As Discussed in Section 3 of [RFC5623], there are three models for inter-layer path computation. They are single PCE computation, multiple PCE with inter-PCE communication and multiple PCE without inter-PCE communication, respectively. For the single PCE computation, the process would be similar to that of the use case in Section 2.1. Thus, this model is not discussed further.

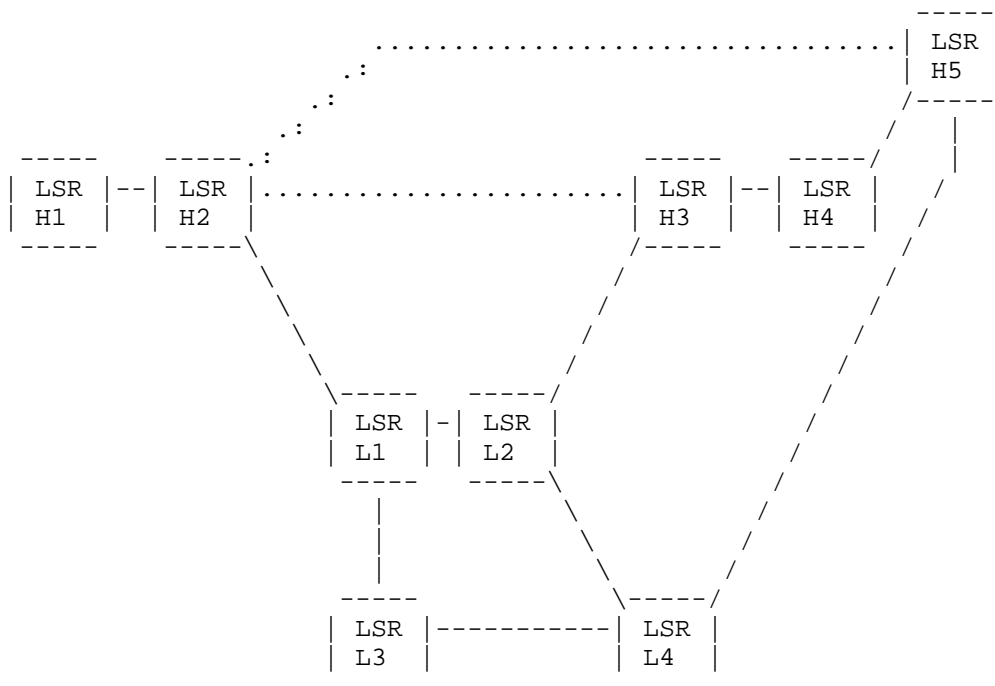


Figure 2: A Two-layer Network Example

In this example, assume a LSP (LSP1: H2-H3) has been established already. A new request comes at H2 to establish a new LSP (LSP2:

from H2 to H5), given the constraint it can share resource with LSP1. This requirement is possible if only one of the LSPs needs to be active and resource sharing is the target.

If multiple PCE with inter-PCE communication model is employed, the path computation request sent by H2 to PCE Hi will be passed to PCE Lo since there is no resource readily available in the upper layer. So it leaves to the PCE Lo to compute a path in the lower layer in order to support the upper layer request. In this case, PCE Lo is required to compute a path between H2 and H5 under the constraint that it can share the resource with that of the LSP1. Assume here LSP1 goes from H2, via L1-L2 to H3. So when PCE Lo computes the path for LSP2, it can view the resource used by LSP1 available. For example, PCE Lo may choose H2-L1-L2-L4-H5 as the computation result.

The issue to solve during this procedure is that PCE Hi can only use LSP1 information (such as its five-tuple LSP information) as the information, how PCE Lo can resolve this information to the actual resource usage in its own layer, i.e. lower layer. This could be solved by edge LSR L1 reporting this higher-lower layer LSP correlation to the Lo PCE as part of the LSP information during the LSP state synchronization process. If needed, it can be later updated when there is a change in this information. Alternatively, the PCE Lo can get this information from other sources, such as network management system, where this information should be stored.

If multiple PCE without inter-PCE communication model is employed, the path computation request in the lower layer will be initiated the border LSR node, i.e., L1. The process would be similar to that of the previous scenario. A point worth noting is that the border LSR node may be able to resolve the higher LSP information itself, such as mapping it to the corresponding LSP in the lower layer, thus PCE Lo do not need to perform this function. Otherwise, the mapping method mentioned above can still be used.

3. Extensions to PCEP

This section provides PCEP extensions. Currently the text focuses only on passive stateful PCE and corresponding PCReq. But if active stateful PCE delegation is used, we would like to convey the same information via RSO in PCRpt. In the passive stateful PCE architecture, a PCC is allowed to specify resource sharing when sending a PCReq message. It also details the processing rule and error codes needed.

3.1. Resource Sharing Object

The PCEP Resource Sharing Object (RSO) is optional. It MAY be carried within a PCRep message so as to indicate the desired resource sharing requirements to be applied by the stateful PCE during path computation.

The RSO object format is compliant with the PCEP object format defined in [RFC5440].

The RSO Object-Class is TBA.

The RSO Object-type is 1.

The format of the RSO object body is:

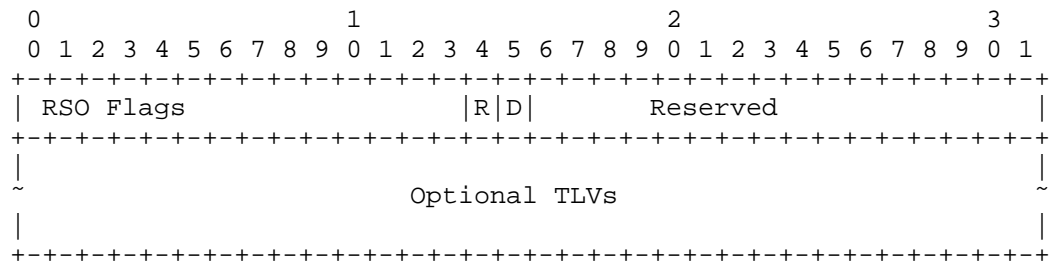


Figure 3: RSO Object Format

RSO codes (16 bits): the objective of the resource sharing. Currently, the following objectives are defined:

D (1 bit): sharing as little as possible.

R (1 bit): sharing as much as possible

It is possible that multiple computation results satisfy the request. Among these results, D set to 1 will select the most separate one, while R set to 1 will select the most sharing one. Both D and R set to 0 don't specify any constraint and will result in a random selection among these results. The combination of D=1 and R=1 is not allowed.

Reserved (2 bytes): This field MUST be set to zero on transmission and MUST be ignored on receipt.

Optional TLVs may be needed to indicate the LSP with which the resource is shared. The LSP Info TLV, include the IPv4-LSP-

IDENTIFIERS TLV and IPv6-LSP IDENTIFIERS TLV, are defined in the same way as in [stateful-pce].

3.2. Processing Rules

To request a path allowing sharing resource with one or multiple existing LSPs, a PCC includes a RSO object in the PCReq message.

On receipt of a PCReq message with a RSO object, a stateful PCE MUST proceed as follows:

- If the RSO object is unknown/unsupported, the PCE will follow procedures defined in [RFC5440]. That is, 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 is discarded.
- If TLV(s) present in the RSO object are unknown/unsupported and the P bit is set, the PCE MUST send a PCErr message with error type 3 or 4 (Unknown / Not supported object) and error value 4 (Unrecognized/Unsupported parameter), and the related path computation request MUST be discarded as defined in [RFC5440].
- If the resource sharing information is extracted correctly, the PCE MUST apply the requested resource sharing requirement.

The procedure of setting R and/or D bit follows the rules defined in Section 3.1. The RSO codes may be locally configured on the requesting nodes via external entities, such as a network management system or the entity that impose the resource sharing requirement.

3.3. Carrying RSO in a PCEP Message

The RSO is applied to an individual path computation request and the format of the PCReq message 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>]

[<metric-list>]

[<OF>]

[<RRO>[<BANDWIDTH>]]

[<IRO>]

[<RSO>[<BANDWIDTH>]]

[<LOAD-BALANCING>]

and where:

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

4. Security Considerations

Security of PCEP is discussed in [RFC5440] and [RFC6952]. The extensions in this document do not change the fundamentals of security for PCEP.

However, the introduction of the RSO provides a vector that may be used to probe for information from a network. For example, a PCC that wants to discover the path of an LSP with which it is not

involved, can issue a PCReq with an RSO and may be able to get back quite a lot of information about the path of the LSP through issuing multiple such requests for different endpoints and analyzing the received results. To protect against this, a PCE should be configured with access and authorization controls such that only authorized PCCs (for example, those within the network) can make computation requests, only specifically authorized PCCs can make requests using the RSO, and resource sharing requests relating to specific LSPs are further limited to a select few PCCs. How such access controls and authorization is managed is outside the scope of this document, but it will at the least include Access Control Lists.

Furthermore, a PCC must be aware that setting up an LSP that shares resources with another LSP may be a way of attacking the other LSP, for example by depriving it of the resources it needs to operate correctly. Thus it is important that, both in PCEP and the associated signaling protocols, only authorized resource sharing is allowed.

5. IANA Considerations

5.1. New Object Type

IANA manages the PCEP Objects code point registry (see [RFC5440]). This is maintained as the "PCEP Objects" sub-registry of the "Path Computation Element Protocol (PCEP) Numbers" registry.

This document defines a new PCEP object, the RSO object, to be carried in PCReq messages. IANA is requested to make the following allocation in the "PCEP Objects" sub-registry:

Object Class	Name	Object Type	Name	Reference

TBA	RSO		Resource Sharing	[this document]

5.2 RSO codes

IANA is requested to create and maintain a new sub-registry named "RSO codes". The following codes are defined in this document:

Bit	Code	Name	Meaning	Reference
0	D		sharing as much as possible	

[this document]

1 R sharing as little as possible

[this document]

6. References

6.1. 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.
- [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.

6.2. Informative References

- [RFC4428] Papadimitriou, D., Mannie., E., ''Analysis of Generalized Multi-Protocol Label Switching (GMPLS)-based Recovery Mechanisms (including Protection and Restoration)'', RFC4428, March 2006.
- [RFC5623] Oki., E., Takeda, T., Le Roux, JL., Farrel, A., ''Framework for PCE-Based Inter-Layer MPLS and GMPLS Traffic Engineering'', RFC5623, September 2009.
- [RFC6952] Jethanandani, M., Patel, K., Zheng, L., ''Analysis of BGP, LDP, PCEP, and MSDP Issues According to the Keying and Authentication for Routing Protocols (KARP) Design Guide'', RFC6952, May 2013.

7. Authors' Addresses

Xian Zhang
Huawei Technologies

Email: zhang.xian@huawei.com

Haomian Zheng
Huawei Technologies

Email: zhenghaomian@huawei.com

Oscar Gonzalez de Dios
Telefonica I+D
Don Ramon de la Cruz 82-84
Madrid 28045
Spain
EMail: ogondio@tid.es

Victor Lopez
Telefonica I+D
Don Ramon de la Cruz 82-84
Madrid 28045
Spain
EMail: vlopez@tid.es

Contributor's Address :

Dhruv Dhody
Huawei Technologies

Email: dhruv.dhody@huawei.com

Igor Bryskin
ADVA Optical

Email: IBryskin@advaoptical.com

PCE Working Group
Internet-Draft
Intended status: Standards Track
Expires: January 5, 2015

Quintin Zhao
Katherine Zhao
Robin Li
Huawei Technologies
Zekun Ke
Tencent Holdings Ltd.
July 4, 2014

The Use Cases for Using PCE as the Central Controller(PCECC) of LSPs
draft-zhao-pce-central-controller-user-cases-01

Abstract

In certain 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/managed 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 use cases for using the PCE as the central controller where LSPs are calculated/setup/initiated/downloaded/maintained through extending the current 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 January 5, 2015.

Copyright Notice

Copyright (c) 2014 IETF Trust and the persons identified as the

document authors. All rights reserved.

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

Table of Contents

1. Introduction	3
1.1. Background	3
1.2. Using the PCE as the Central Controller (PCECC) Approach	4
2. Terminology	7
3. PCEP Requirements	7
4. Use Cases of PCECC for Label Resource Reservations	8
5. Using PCECC for SR without the IGP Extension	9
5.1. Use Cases of PCECC for SR Best Effort(BE) Path	10
5.2. Use Cases of PCECC for SR Traffic Engineering (TE) Path	11
6. Use Cases of PCECC for TE LSP	12
7. Use Cases of PCECC for Multicast LSPs	14
7.1. Using PCECC for P2MP/MP2MP LSPs' Setup	14
7.2. Use Cases of PCECC for the Resiliency of P2MP/MP2MP LSPs	15
7.2.1. PCECC for the End-to-End Protection of the P2MP/MP2MP LSPs	15
7.2.2. PCECC for the Local Protection of the P2MP/MP2MP LSPs	16
8. Use Cases of PCECC for LSP in the Network Migration	17
9. The Considerations for PCECC Procedure and PCEP extensions	19
10. IANA Considerations	19
11. Security Considerations	19
12. Acknowledgments	19
13. References	19
13.1. Normative References	19
13.2. Informative References	19

1. Introduction

1.1. Background

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.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.ali-pce-remote-initiated-gmpls-lsp] 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-spring-segment-routing] provides an introduction to SR technology. The corresponding IS-IS and OSPF extensions are specified in [I-D.ietf-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.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. These key MPLS features include MPLS P2P LSP, P2MP/MP2MP LSP, MPLS protection mechanism etc. 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.

1.2. 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.

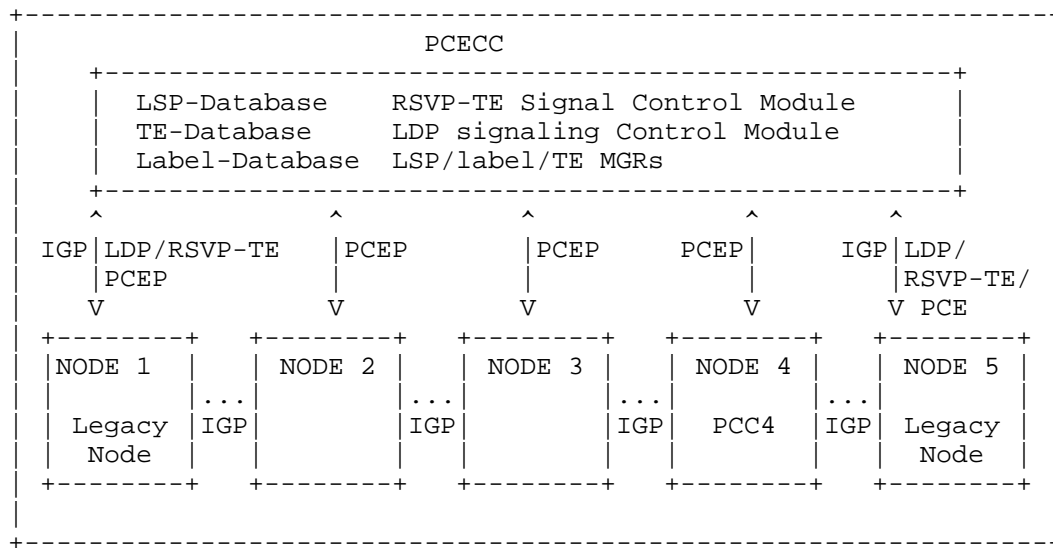


Figure 1: PCECC Architecture

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.

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-mpls-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. The solution choices are described in [I-D.li-mpls-global-label-framework].

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.

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.

For the centralized network, the performance achieved through distributed system can not be easy matched if all of the forwarding path is computed, downloaded and maintained by the centralized controller. The performance can be improved by supporting part of the forwarding path in the PCECC network through the segment routing mechanism except that the adjacency IDs for all the network nodes and links are propagated through the centralized controller instead of using the IGP extension.

The node and link adjacency IDs can be negotiated through the PCECC with each PCECC clients and these IDs can be just taken from the global label range which has been negotiated already.

With the capability of supporting SR within the PCECC architecture, all the p2p forwarding path protection use cases described in the draft [I-D.ietf-spring-resiliency-use-cases] will be supported too within the PCECC network. These protection alternatives include end-to-end path protection, local protection without operator management and local protection with operator management.

With the capability of global label and local label existing at the same time in the PCECC network, PCECC will use compute, setup and maintain the P2MP and MP2MP lsp using the local label range for each network nodes.

With the capability of setting up/maintaining the P2MP/MP2MP LSP within the PCECC network, it is easy to provide the end-end managed path protection service and the local protection with the operation management in the PCECC network for the P2MP/MP2MP LSP, which includes both the RSVP-TE P2MP based LSP and also the mLDP based LSP.

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. Use Cases of PCECC for Label Resource Reservations

Example 1 to 2 are based on network configurations illustrated using the following figure:

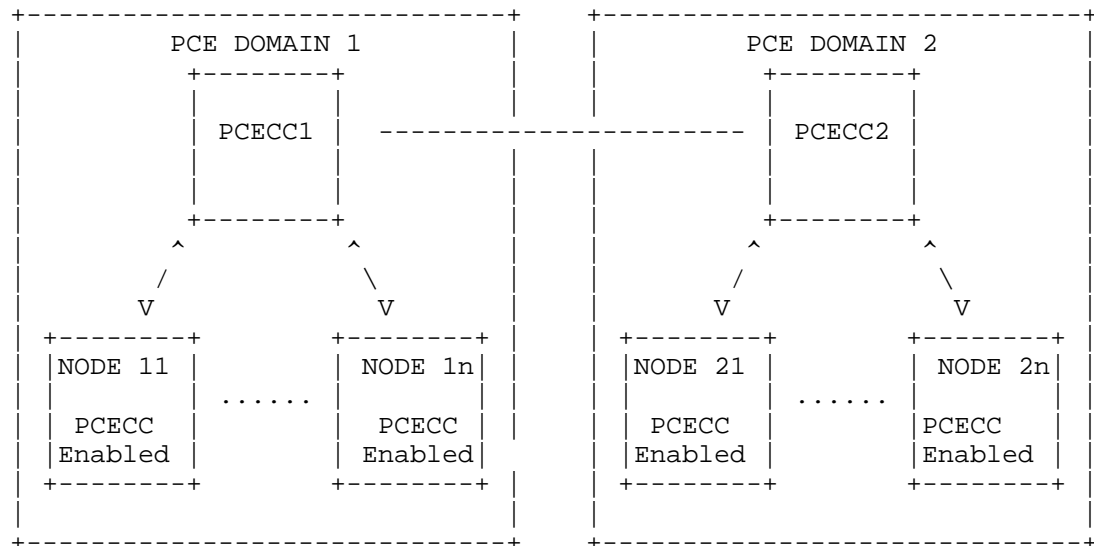


Figure 2: Using PCECC for Global Label Allocation

Example 1: Shared Global Label Range Reservation

- o PCECC Clients nodes report MPLS label capability to the central controller PCECC.
- o The central controller PCECC collects MPLS label capability of all nodes. Then PCECC can calculate the shared MPLS global label range for all the PCECC client nodes.
- o In the case that the shared global label range need to be negotiated across multiple domains, the central controllers of these domains need to be communicate to negotiate a common global label range.
- o The central controller PCECC notifies the shared global label range to all PCECC client nodes.

Example 2: Global Label Allocation

- o PCECC Client node1 send global label allocation request to the central controller PCECC1.
- o The central controller PCECC1 allocates the global label for FEC1 from the shared global label range and sends the reply to the client node1.
- o The central controller PCECC1 notifies the allocated label for FEC1 to all PCECC client nodes within domain 1.

5. Using PCECC for SR without the IGP Extension

For the centralized network, the performance achieved through distributed system can not be easily matched if all of the forwarding path is computed, downloaded and maintained by the centralized controller. The performance can be improved by supporting part of the forwarding path in the PCECC network through the segment routing mechanism except that node segment IDs and adjacency segment IDs for all the network are allocated dynamically and propagated through the centralized controller instead of using the IGP extension.

When the PCECC is used for the distribution of the node segment ID and adjacency segment ID, the node segment ID is allocated from the global label pool. For the allocation of adjacency segment ID, there are two choices, the first choice is that it is allocated from the local label pool, the second choice is that it is allocated from the global label pool. The advantage for the second choice is that the depth of the label stack for the forwarding path encoding will be reduced since adjacency segment ID can signal the forwarding path without adding the node segment ID in front of it. In this version of the draft, we use the first choice for now. We may update the draft to reflect the use of the second choice.

Same as the SR solutions, when PCECC is used as the central controller, the support of FRR on any topology can be pre-computed and setup without any additional signaling (other than the regular IGP/BGP protocols) including the support of shared risk constraints, support of node and link protection and support of microloop avoidance.

The following example illustrates the use case where the node segment ID and adjacency segment ID are allocated from the global label allocated for SR path.

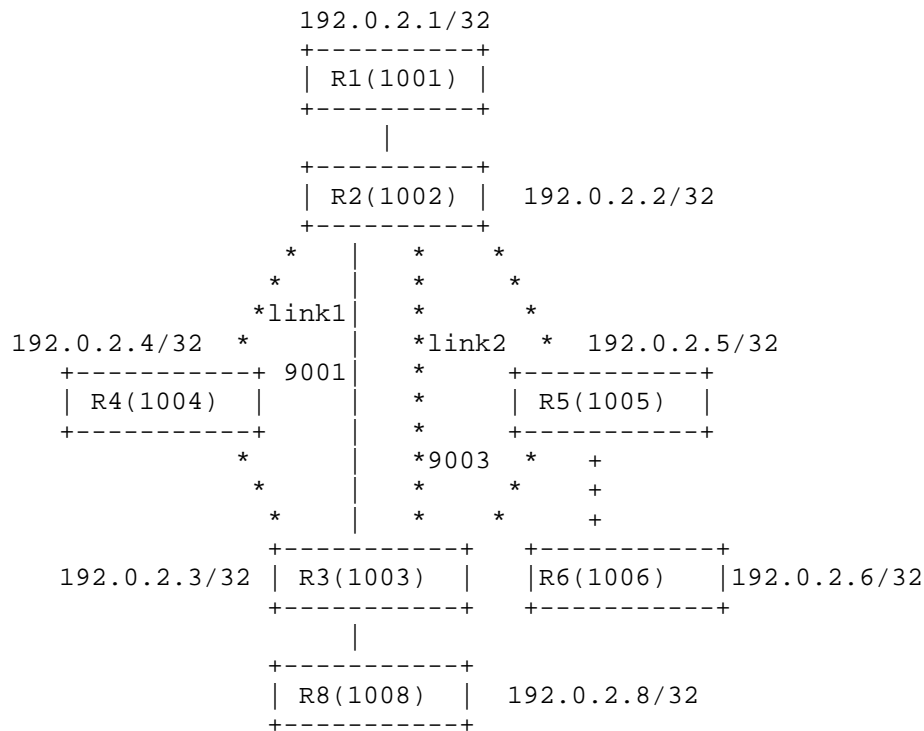


Figure 3: Using PCECC for SR Path

5.1. Use Cases of PCECC for SR Best Effort(BE) Path

In this mode of the solution, the PCECC just need to allocate the node segment ID and adjacency ID without calculating the explicit path for the SR path. The ingress of the forwarding path just need to encapsulate the destination node segment ID on top of the packet. All the intermediate nodes will forward the packet based on the final destination node segment id. It is similar to the LDP LSP forwarding except that label swapping is using the same global label both for the in segment and out segment in each hop.

The p2p SR BE path examples are explained as bellow:

Note that the node segment id for each node from the shared global labels ranges negotiated already.

Example 1:

R1 may send a packet to R8 simply by pushing an SR header with segment list {1008}. The path can be: R1-R2-R3-R8 or R1-R2-R5-R8

depending on the route calculation on node R2.

Example 2: local link/node protection:

For the packet which has destination of R3 and after that, R2 may preinstalled the backup forwarding entry to protect the R4 node, the pre-installed the backup path can go through either node5 or link1 or link2 between R2 and R3. The backup path calculation is locally decided by R2 and any existing IP FRR algorithms can be used here.

5.2. Use Cases of PCECC for SR Traffic Engineering (TE) Path

In the case of traffic engineering path is needed, the PCECC need to allocate the node segment ID and adjacency ID, and at the same time PCECC calculates the explicit path for the SR path and pass this explicit path represented with a sequence of node segment id and adjacency id. The ingress of the forwarding path need to encapsulate the stack of node segment id and adjacency id on top of the packet. For the case where strict traffic engineering path is needed, all the intermediate nodes and links will be specified through the stack of labels so that the packet is forwarded exactly as it is wanted.

Even though it is similar to TE LSP forwarding where forwarding path is engineered, but the Qos is only guaranteed through the enforce of the bandwidth admission control. As for the RSVP-TE LSP case, Qos is guaranteed through the link bandwidth reservation in each hop of the forwarding path.

The p2p SR traffic engineering path examples are explained as bellow:

Note that the node segment id for each node is allocated from the shared global labels ranges negotiated already and adjacency segment ids for each link are allocated from the local label pool for each node.

Example 1:

R1 may send a packet P1 to R8 simply by pushing an SR header with segment list {1008}. The path should be: R1-R2-R3-R8.

Example 2:

R1 may send a packet P2 to R8 by pushing an SR header with segment list {1002, 9001, 1008}. The path should be: R1-R2-(1)link-R3-R8.

Example 3:

R1 may send a packet P3 to R8 while avoiding the links between R2 and

R3 by pushing an SR header with segment list {1004, 1008}. The path should be : R1-R2-R4-R3-R8

The p2p local protection examples for SR TE path are explained as below:

Example 4: local link protection:

- o R1 may send a packet P4 to R8 by pushing an SR header with segment list {1002, 9001, 1008}. The path should be: R1-R2-(1)link-R3-R8.
- o When node R2 receives the packet from R1 which has the header of R2- (1)link-R3-R8, and also find out there is a link failure of link1, then it will send out the packet with header of R3-R8 through link2.

Example 5: local node protection:

- o R1 may send a packet P5 to R8 by pushing an SR header with segment list {1004, 1008}. The path should be : R1-R2-R4-R3-R8.
- o When node R2 receives the packet from R1 which has the header of {1004, 1008}, and also find out there is a node failure for node4, then it will send out the packet with header of {1005, 1008} to node5 instead of node4.

6. Use Cases of PCECC for TE LSP

In the previous sections, we have discussed the cases where the SR path is setup through the PCECC. Although those cases give the simplicity and scalability, but there are existing functionalities for the traffic engineering path such as the bandwidth guarantee through the full forwarding path and the multicast forwarding path which SR based solution cannot solve. Also there are cases where the depth of the label stack may have been an issue for existing deployment and certain vendors.

So to address these issues, PCECC architecture should also support the TE LSP and multicast LSP functionalities. To achieve this, the existing PCEP can be used to communicate between the PCE server and PCE's client PCC for exchanging the path request and reply information regarding to the TE LSP info. In this case, the TE LSP info is not only the path info itself, but it includes the full forwarding info. Instead of letting the ingress of LSP to initiate the LSP setup through the RSVP-TE signaling protocol, with minor extensions, we can use the PCEP to download the complete TE LSP forwarding entries for each node in the network.

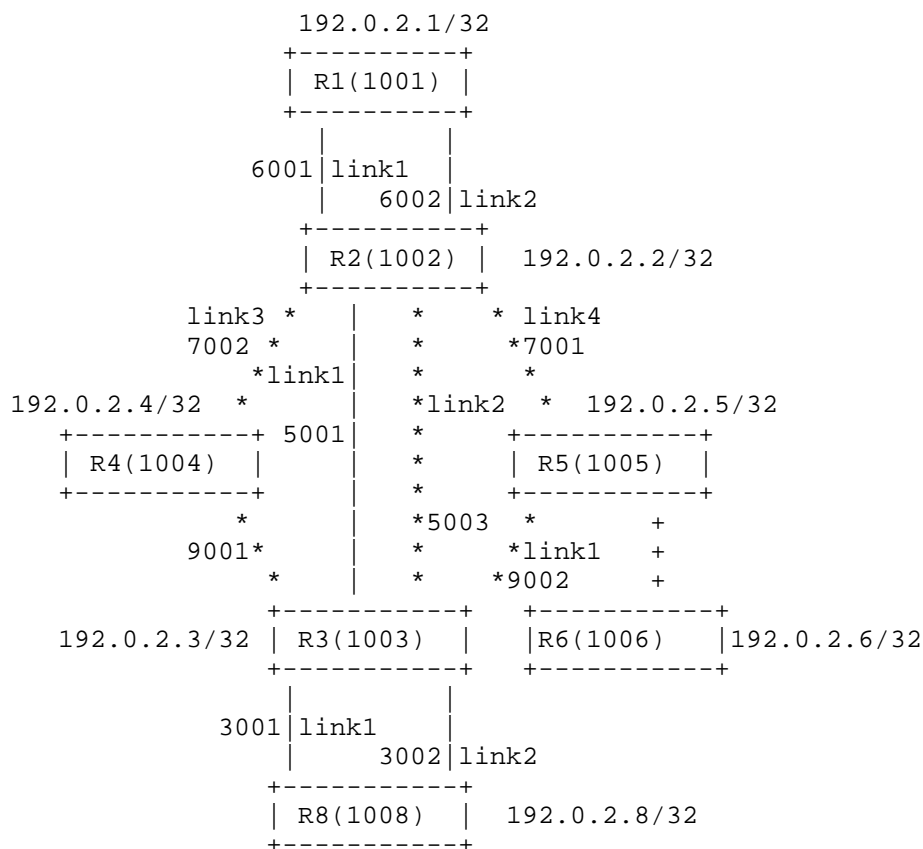


Figure 4: Using PCECC for TE LSP

TE LSP Setup Example

- o Node1 sends a path request message for the setup of TE LSP from R1 to R8.
- o PCECC program each node along the path from R1 to R8 with the primary path: {R1, link1, 6001}, {R2, link3, 7002}, {R4, link0, 9001}, {R3, link1, 3001}, {R8}.
- o For the end to end protection, PCECC program each node along the path from R1 to R8 with the secondary path: {R1, link2, 6002}, {R2, link4, 7001}, {R5, link1, 9002}, {R3, link2, 3002}, {R8}.
- o It is also possible to have a secondary backup path for the local node protection setup by PCECC. For example GBP[not] the primary path is still same as what we have setup so far, then to protect

the node R4 locally, PCECC can program the secondary path like this: {R1, link1, 6001}, {R2, link1, 5001}, {R3, link1, 3001}, {R8}. By doing this, the node R4 is locally protected.

7. Use Cases of PCECC for Multicast LSPs

The current multicast LSPs are setup either using the RSVP-TE P2MP or mLDP protocols. The setup of these LSPs not only need a lot of manual configurations, but also it is also complex when the protection is considered. By using the PCECC solution, the multicast LSP can be computed and setup through centralized controller which has the full picture of the topology and bandwidth usage for each link. It not only reduces the complex configurations comparing the distributed RSVP-TE P2MP or mLDP signal lings, but also it can compute the disjoint primary path and secondary path efficiently.

7.1. Using PCECC for P2MP/MP2MP LSPs' Setup

With the capability of global label and local label existing at the same time in the PCECC network, PCECC will use compute, setup and maintain the P2MP and MP2MP lsp using the local label range for each network nodes.

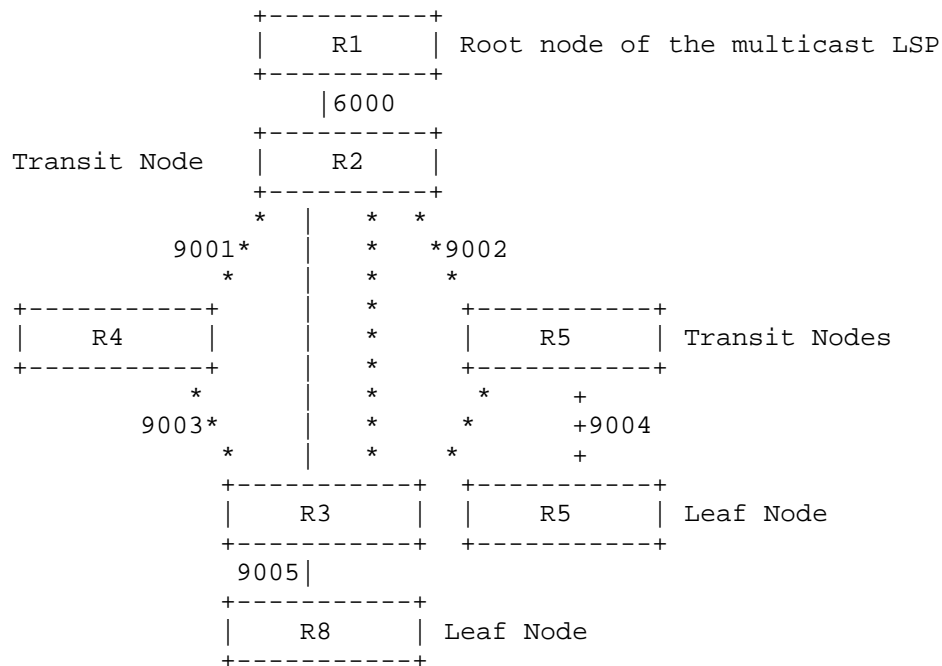


Figure 5: Using PCECC for P2MP TE LSP

The P2MP examples are explained here:

Step1: R1 may send a packet P1 to R2 simply by pushing an label of 6000 to the packet.

Step2: After R2 receives the packet with label 6000, it will forwarding to R4 by pushing header of 9001 and R5 by pushing header of 9002.

Step3: After R4 receives the packet with label 9001, it will forwarding to R3 by pushing header of 9003. After R5 receives the packet with label 9002, it will forwarding to R5 by pushing header of 9004.

Step3: After R3 receives the packet with label 9003, it will forwarding to R8 by pushing header of 9005

7.2. Use Cases of PCECC for the Resiliency of P2MP/MP2MP LSPs

7.2.1. PCECC for the End-to-End Protection of the P2MP/MP2MP LSPs

In this section we describe the end-end managed path protection service and the local protection with the operation management in the PCECC network for the P2MP/MP2MP LSP, which includes both the RSVP-TE P2MP based LSP and also the mLDP based LSP.

An end-to-end protection (for nodes and links) principle can be applied for computing backup P2MP or MP2MP LSPs. During computation of the primarily multicast trees, PCECC server may also be taken into consideration to compute a secondary tree. A PCE may compute the primary and backup P2MP or MP2MP LSP together or sequentially.

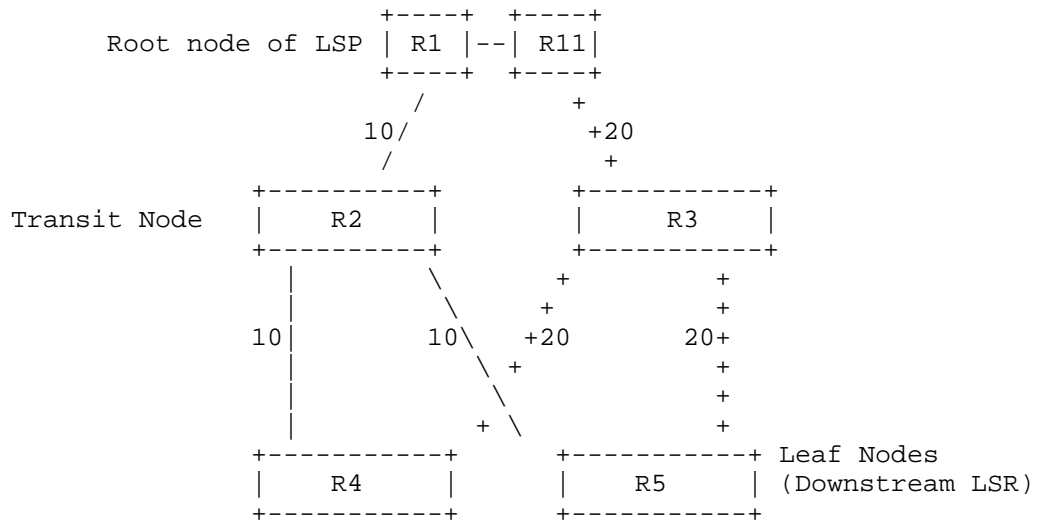


Figure 6: Using PCECC for P2MP TE End-to-End Protection

In the example above, when the PCECC setup the primary multicast tree from the root node R1 to the leafs, which is R1->R2->{R4, R5}, at same time, it can setup the backup tree, which is R11->R3->{R4, R5}. Both the these two primary forwarding tree and secondary forwarding tree will be downloaded to each routers along the primary path and the secondary path. The traffic will be forwarded through the R1->R2->{R4, R5} path normally, and when there is a node in the primary tree, then the root node R1 will switch the flow to the backup tree, which is R11->R3->{R4, R5}. By using the PCECC, the path computation and forwarding path downloading can all be done without the complex signaling used in the P2MP RSVP-TE or mLDP.

7.2.2. PCECC for the Local Protection of the P2MP/MP2MP LSPs

In this section we describe the local protection service in the PCECC network for the P2MP/MP2MP LSP.

While the PCECC sets up the primary multicast tree, it can also build the back LSP among PLR, the protected node, and MPs (the downstream nodes of the protected node). In the cases where the amount of downstream nodes are huge, this mechanism can avoid unnecessary packet duplication on PLR, so that protect the network from traffic congestion risk.

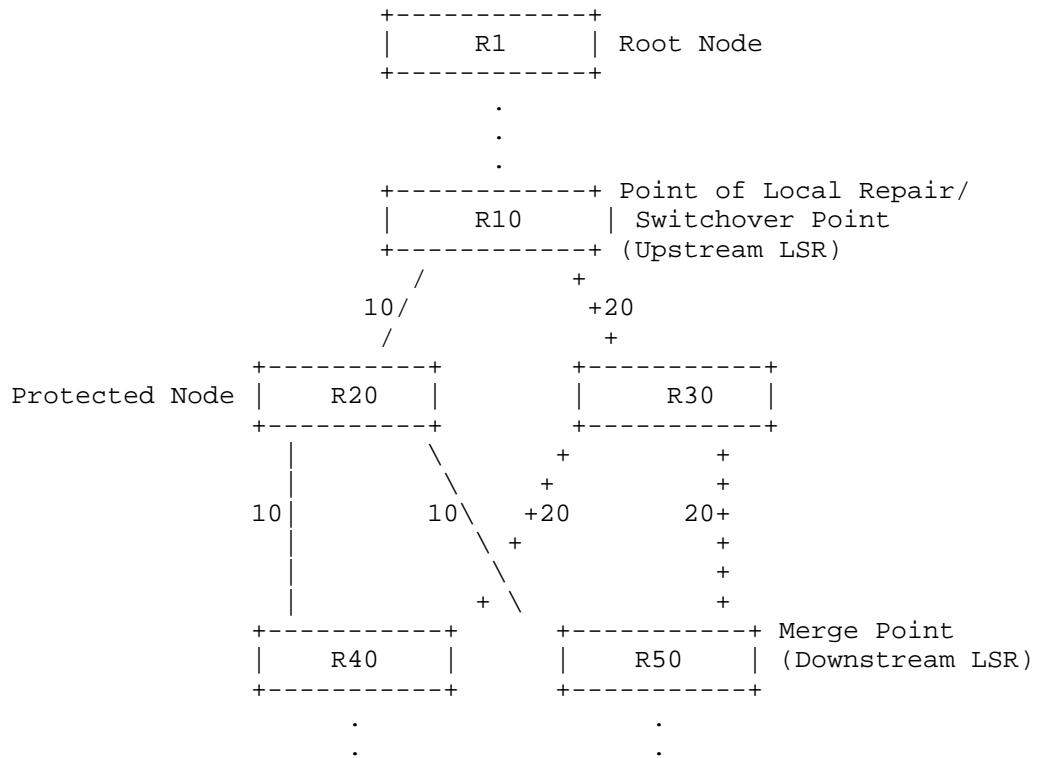


Figure 7: Using PCECC for P2MP TE LocalProtection

In the example above, when the PCECC setup the primary multicast path around the PLR node R10 to protect node R20, which is R10->R20->{R40, R50}, at same time, it can setup the backup path R10->R30->{R40, R50}. Both the these two primary forwarding path and secondary forwarding path will be downloaded to each routers along the primary path and the secondary path. The traffic will be forwarded through the R10->R20->{R40, R50} path normally, and when there is a node failure for node R20, then the PLR node R10 will switch the flow to the backup path, which is R10->R30->{R40, R50}. By using the PCECC, the path computation and forwarding path downloading can all be done without the complex signaling used in the P2MP RSVP-TE or mLDP.

8. Use Cases of PCECC for LSP in the Network Migration

One of the main advantages for PCECC solution is that it has backward compatibility naturally since the PCE server itself can function as a proxy node of MPLS network for all the new nodes which don't support the existing MPLS signaling protocol anymore.

As it is illustrated in the following example, the current network will migrate to a total PCECC controlled network gradually by replacing the legacy nodes. During the migration, the legacy nodes still need to signal using the existing MPLS protocol such as LDP and RSVP-TE, and the new nodes setup their portion of the forwarding path through PCECC directly. With the PCECC function as the proxy of these new nodes, MPLS signaling can populate through network as normal.

Example described in this section is based on network configurations illustrated using the following figure:

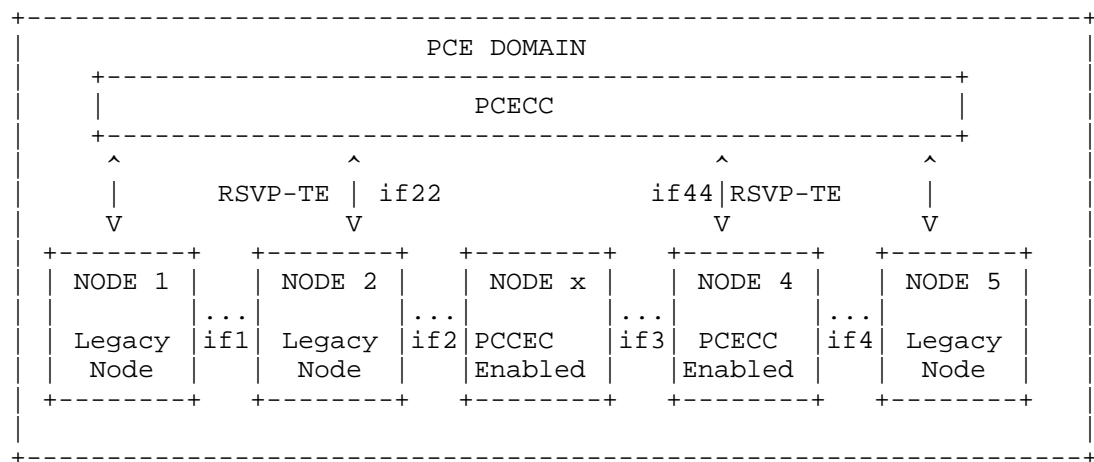


Figure 8: Using PCECC During Migration

Example: PCECC Initiated LSP Setup In the Network Migration

In this example, there are five nodes for the TE LSP from head end (node1) to the tail end (node5). Where the NodeX is central controlled and other nodes are legacy nodes.

- o Node1 sends a path request message for the setup of LSP destinating to Node5.
- o PCECC sends a reply message for LSP setup with path (node1, if1), (node2, if22), (node-PCECC, if44), (node4, if4), Nnode5.
- o Node1, Node2, Node-PCECC, Node 5 will setup the LSP to Node5 normally using the local label as normal.

- o Then the PCECC will program the outsegment of Node2, the insegment of Node4, and the insegment/outsegment for NodeX.

9. The Considerations for PCECC Procedure and PCEP extensions

The PCECC's procedures and PCEP extensions is defined in [I-D.zhao-pce-pcep-extension-for-pce-controller].

10. IANA Considerations

This document does not require any action from IANA.

11. Security Considerations

TBD.

12. Acknowledgments

We would like to thank Robert Tao, Changjiang Yan, Tieying Huang 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.

[RFC5440]

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

13.2. 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, J.L.,
Vasseur, J.P., and
Y. Lee, "Encoding
of Objective
Functions in the
Path Computation
Element
Communication
Protocol (PCEP)",
RFC 5541,
June 2009.

[I-D.filsfils-spring-segment-routing]

Filsfils, C.,
Previdi, S.,
Bashandy, A.,
Decraene, B.,
Litkowski, S.,
Horneffer, M.,
Milojevic, I.,
Shakir, R., Ytti,
S., Henderickx, W.,
Tantsura, J., and
E. Crabbe, "Segment
Routing
Architecture", draf
t-filsfils-spring-
segment-routing-04
(work in progress),
July 2014.

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

Crabbe, E., Minei,
I., Medved, J., and
R. Varga, "PCEP
Extensions for
Stateful PCE", draf
t-ietf-pce-
stateful-pce-09
(work in progress),

June 2014.

[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., Dios, O., and X. Zhang, "Path Computation Element Communication Protocol (PCEP) Extensions for remote-initiated GMPLS LSP Setup", draft-ali-pce-remote-initiated-gmpls-lsp-03 (work in progress), February 2014.

[I-D.ietf-isis-segment-routing-extensions]

Previdi, S., Filsfils, C., Bashandy, A., Gredler, H., Litkowski, S., Decraene, B., and J. Tantsura, "IS-IS Extensions for Segment Routing", draft-ietf-isis-segment-routing-extensions-02 (work in progress), June 2014.

[I-D.psenak-ospf-segment-routing-extensions]

Psenak, P.,
Previdi, S.,

- Filsfils, C.,
Gredler, H.,
Shakir, R.,
Henderickx, W., and
J. Tantsura, "OSPF
Extensions for
Segment Routing", d
raft-psenak-ospf-
segment-routing-
extensions-05 (work
in progress),
June 2014.
- [I-D.sivabalan-pce-segment-routing] Sivabalan, S.,
Medved, J.,
Filsfils, C.,
Crabbe, E., and R.
Raszuk, "PCEP
Extensions for
Segment Routing", d
raft-sivabalan-pce-
segment-routing-02
(work in progress),
October 2013.
- [I-D.li-mpls-global-label-usecases] Li, Z., Zhao, Q.,
Yang, T., and R.
Raszuk, "Use Cases
of MPLS Global
Label", draft-li-
mpls-global-label-
usecases-02 (work
in progress),
July 2014.
- [I-D.li-mpls-global-label-framework] Li, Z., Zhao, Q.,
Chen, X., Yang, T.,
and R. Raszuk, "A
Framework of MPLS
Global Label", draf
t-li-mpls-global-
label-framework-02
(work in progress),
July 2014.
- [I-D.zhao-pce-pcep-extension-for-pce-controller] Zhao, Q., Zhao, K.,
Dhody, D., and B.
Zhang, "PCEP

Procedures and
Protocol Extensions
for Using PCE as a
Central Controller
(PCECC) of LSPs", d
raft-zhao-pce-pcep-
extension-for-pce-
controller-00 (work
in progress),
February 2014.

[I-D.ietf-spring-resiliency-use-cases]

Francois, P.,
Filsfils, C.,
Decraene, B., and
R. Shakir, "Use-
cases for
Resiliency in
SPRING", draft-
ietf-spring-
resiliency-use-
cases-00 (work in
progress),
May 2014.

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

