

PCE Working Group
Internet-Draft
Intended status: Standards Track
Expires: January 7, 2016

C. Barth
R. Torvi
Juniper Networks
P. Bedard
Cox Communications
July 6, 2015

PCEP Extensions for RSVP-TE Local-Protection with PCE-Stateful
draft-cbrt-pce-stateful-local-protection-00

Abstract

Stateful PCE [ietf-pce-stateful-pce] can apply global concurrent optimizations to optimize LSP placement. In a deployment where a PCE is used to compute all the paths, it may be beneficial for the local protection paths to also be computed by the PCE. This document defines extensions needed for the setup and management of RSVP-TE protection paths by the PCE.

Status of This Memo

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

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

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

This Internet-Draft will expire on January 7, 2016.

Copyright Notice

Copyright (c) 2015 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. Local Protection Overview	3
4. Extensions for the LSPA object	4
4.1. The Weight TLV	4
4.2. The Bypass TLV	4
4.3. The LOCALLY-PROTECTED-LSPS TLV	5
5. IANA considerations	7
5.1. PCEP-Error Object	7
5.2. PCEP TLV Type Indicators	7
6. Security Considerations	7
7. Acknowledgements	7
8. References	7
8.1. Normative References	7
8.2. Informative References	8
Appendix A. Additional Stuff	9
Authors' Addresses	9

1. Introduction

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

Stateful PCE [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 and allow delegation of control of LSPs to PCEs.

In a network where all LSPs have control delegated to a PCE, the PCE can apply global concurrent optimization to optimize LSP placement. The PCE can also control the timing and sequence of path computation and applying path changes. In a deployment where a PCE is used to compute all the paths, it may be beneficial for the protection paths to also be controlled through the PCE. This document defines extensions needed for the setup and management of protection paths by the PCE.

Benefits of stateful synchrhonization and control of the protection paths include:

- o Better control over traffic after a failure and more deterministic path computation of protection paths. The PCE can optimize the protection path based on data not available to the PCC, for instance the PCE can make sure the protection path will not violate the delay specified by [I-D.ietf-pce-pcep-service-aware].
- o Satisfy more complex constraints and diversity requirements, such as maintaining diverse paths for LSPs as well as their local protection paths.
- o Given the PCE's global view of network resources, act as a form of LSP admission control into a protection path to ensure links are not overloaded during failure events.
- o On a PLR with multiple available protection routes, allows the PCE to map LSPs to all available protection routes versus a single best protection route.
- o Most of the benefits stated in the stateful PCE applicability draft [I-D.ietf-pce-stateful-pce-app-04] apply equally to protection paths.

2. Terminology

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

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

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

3. Architectural Overview

3.1. Local Protection Overview

Local protection refers to the ability to locally route around failure of an LSP. Two types of local protection are possible:

- (1) 1:1 protection - the protection path protects a single LSP.
- (2) N:1 protection - the protection path protects multiple LSPs traversing the protected resource.

It is assumed that the PCE knows what resources require protection through mechanisms outside the scope of this document. In a PCE controlled deployment, support of 1:1 protection has limited applicability, and can be achieved as a degenerate case of 1:N protection. For this reason, local protection will be discussed only for the N:1 case.

Local protection requires the setup of a bypass at the PLR. This bypass can be PCC-initiated and delegated, or PCE-initiated. In either case, the PLR MUST maintain a PCEP session to the PCE. A bypass identifier (the name of the bypass) is required for disambiguation as multiple bypasses are possible at the PLR. There are two types of Bypass LSP mappings:

(1) Independent Bypass LSP Mapping: In this case Bypass LSP mapping is handled by a local policy on PCC and the PCC reports all mappings to the PCE. In other words, bypass LSP(s) are mapped to any protected LSP(s) that satisfy PCC local policy.

(2) Dependent Bypass LSP mapping: Mapping of LSPs to bypass is done through a new TLV, the LOCALLY-PROTECTED-LSPS TLV in the LSP Update message from PCE to PLR. See section 4.3. When an LSP requiring protection is set up through the PLR, the PLR checks if it has a mapping to a bypass and only provides protection if such a mapping exists. The status of bypasses and what LSPs are protected by them is communicated to the PCE via LSP Status Report messages.

4. Extensions for the LSPA object

4.1. The Weight TLV

This TLV will be discussed in a future version of this document.

4.2. The Bypass TLV

The facility backup method creates a bypass tunnel to protect a potential failure point. The bypass tunnel protects a set of LSPs with similar backup constraints [RFC4090].

A PCC can delegate a bypass tunnel to PCE control or a PCE can provision the bypass tunnel via a PCC. The procedures for bypass instantiation rely on the extensions defined in PCE-Initiated LSP [ietf-pce-pce-initiated-lsp] and will be detailed in a future version of this document.

The Bypass TLV carries information about the bypass tunnel. It is included in the LSPA Object in LSP State Report and LSP Update Request messages.

The format of the IPv4 Bypass TLV is shown in the following figure:

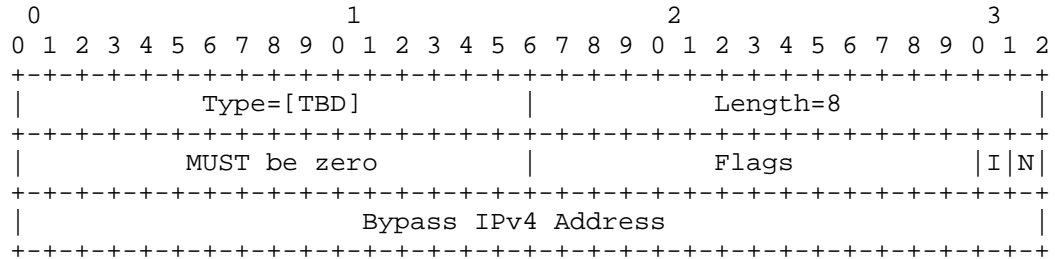


Figure 1: IPv4 Bypass TLV format

The type of the TLV is [TBD] and it has a fixed length of 8 octets. The value contains the following fields:

Flags (16 bit)

N (Node Protection - 1 bit): The N flag indicates whether the Bypass is used for node-protection. If the N flag is set to 1, the Bypass is used for node-protection. If the N flag is 0, the Bypass is used for link-protection.

I (Local Protection In Use - 1 bit): The I Flag indicates that local repair mechanism is in use.

Bypass IPv4 address: For link protection, the Bypass IPv4 Address is the nexthop address of the protected link in the paths of the protected LSPs. For node protection, the Bypass IPv4 Address is the node addresses of the protected node.

If the Bypass TLV is included, then the LSPA object MUST also carry the SYMBOLIC-PATH-NAME TLV as one of the optional TLVs. Failure to include the mandatory SYMBOLIC-PATH-NAME TLV MUST trigger PCerr of type 6 (Mandatory Object missing) and value TBD (SYMBOLIC-PATH-NAME TLV missing for bypass LSP)

4.3. The LOCALLY-PROTECTED-LSPS TLV

The IPV4-LOCALLY-PROTECTED-LSPS TLV in the LSPA Object contains a list of LSPs protected by the bypass tunnel.

The format of the Locally protected LSPs TLV is shown in the following figure:

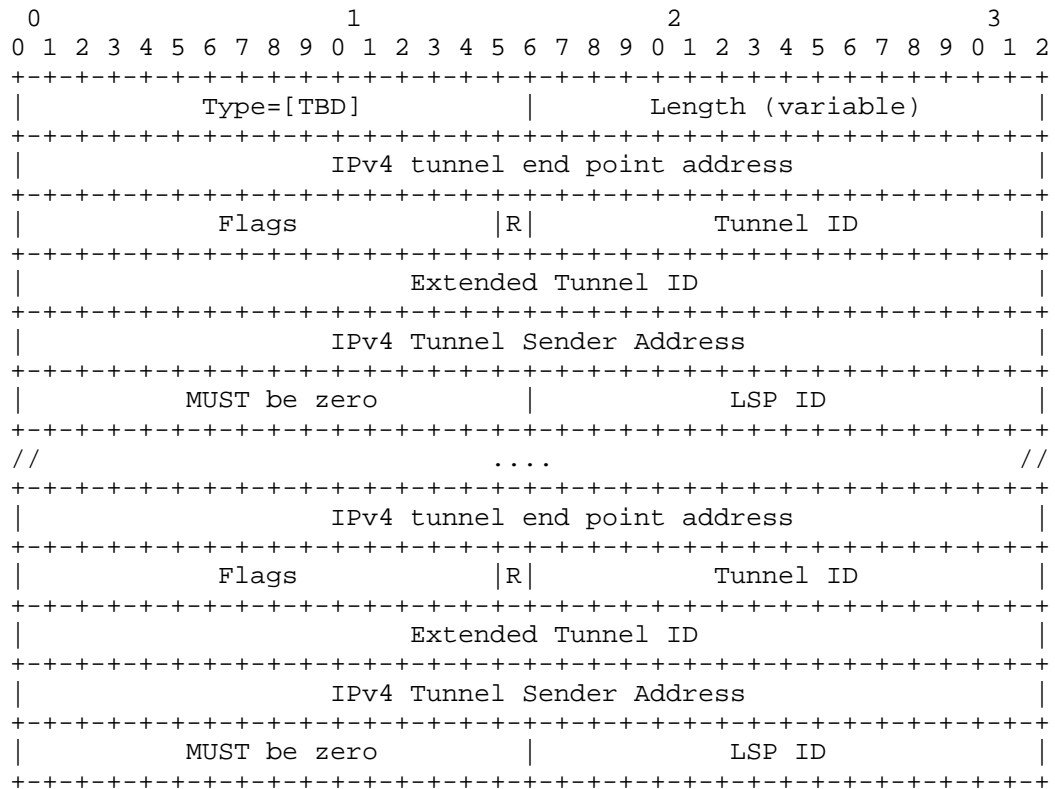


Figure 2: IPv4 Locally protected LSPs TLV format

The type of the TLV is [TBD] and it is of variable length. The value contains one or more LSP descriptors including the following fields filled per [RFC3209]

IPv4 Tunnel end point address: As defined in [RFC3209], Section 4.6.1.1

Flags (16 bit)

R(Remove - 1 bit): The R flag indicates that the LSP has been removed from the list of LSPs protected by the bypass tunnel.

Tunnel ID: As defined in [RFC3209], Section 4.6.1.1

Extended Tunnel ID: As defined in [RFC3209], Section 4.6.2.1

IPv4 Tunnel Sender address: As defined in [RFC3209], Section 4.6.2.1

LSP ID: As defined in RFC 3209

5. IANA considerations

5.1. PCEP-Error Object

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

Error-Type Meaning 6 Mandatory Object missing Error-value=TBD:
 SYMBOLIC-PATH-NAME TLV missing for a path where the S-bit is set in the LSPA object. Error-value=TBD: SYMBOLIC-PATH-NAME TLV missing for a bypass path.

5.2. PCEP TLV Type Indicators

This document defines the following new PCEP TLVs:

Value #	Meaning	Reference
???	Bypass	This Document
???	Weight	This Document
???	LOCALLY-PROTECTED-LSPS	This Document

Table 1: New PCEP TLVs

6. Security Considerations

The same security considerations apply at the PLR as those describe for the head end in PCE Initiated LSPs [ietf-pce-pce-initiated-lsp].

7. Acknowledgements

We would like to thank Ambrose Kwong for his contributions to this document.

8. References

8.1. Normative References

[ietf-pce-pce-initiated-lsp]
 Crabbe, E., Sivabalan, S., and R. Verga, "PCEP Extensions for PCE-initiated LSP Setup in a Stateful PCE Model", 2014.

- [ietf-pce-stateful-pce]
Crabbe, E., Medved, J., Minie, I., and R. Verga, "PCEP Extensions for Stateful PCE", 2015.
- [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", September 1997.
- [RFC3209] Awduche, D., Berger, L., Gan, D., Li, T., Srinivasan, V., and G. Swallow, "RSVP-TE: Extensions to RSVP for LSP Tunnels", December 2001.
- [RFC4090] Pan, P., Swallow, G., and A. Atlas, "Fast Reroute Extensions to RSVP-TE for LSP Tunnels", May 2005.
- [RFC5226] Narten, T. and H. Alvestrand, "Guidelines for Writing an IANA Considerations Section in RFCs", May 2008.
- [RFC5440] Vasseur, JP. and JL. Le Roux, "Path Computation Element (PCE) Communication Protocol (PCEP)", March 2009.

8.2. Informative References

- [I-D.narten-iana-considerations-rfc2434bis]
Narten, T. and H. Alvestrand, "Guidelines for Writing an IANA Considerations Section in RFCs", draft-narten-iana-considerations-rfc2434bis-09 (work in progress), March 2008.
- [RFC2629] Rose, M., "Writing I-Ds and RFCs using XML", RFC 2629, June 1999.
- [RFC3552] Rescorla, E. and B. Korver, "Guidelines for Writing RFC Text on Security Considerations", BCP 72, RFC 3552, July 2003.
- [RFC4655] Farrel, A., Vasseur, J., and J. Ash, "A Path Computation Element (PCE)-Based Architecture", August 2006.
- [RFC4657] Ash, J. and J. Le Roux, "Path Computation Element (PCE) Communication Protocol Generic Requirements", September 2006.

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

Appendix A. Additional Stuff

This becomes an Appendix.

Authors' Addresses

Colby Barth
Juniper Networks
Sunnyvale, CA
USA

Email: cbarth@juniper.net

Raveendra Torvi
Juniper Networks
Sunnyvale, CA
USA

Email: rtorvi@juniper.net

Phil Bedard
Cox Communications
Atlanta, GA
USA

Email: Phil.Bedard@cox.com

PCE Working Group
Internet-Draft
Intended status: Standards Track
Expires: December 30, 2018

C. Barth
R. Torvi
Juniper Networks
June 28, 2018

PCEP Extensions for RSVP-TE Local-Protection with PCE-Stateful
draft-cbrt-pce-stateful-local-protection-01

Abstract

Stateful PCE [RFC8231] can apply global concurrent optimizations to optimize LSP placement. In a deployment where a PCE is used to compute all the paths, it may be beneficial for the local protection paths to also be computed by the PCE. This document defines extensions needed for the setup and management of RSVP-TE protection paths by the PCE.

Status of This Memo

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

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

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

This Internet-Draft will expire on December 30, 2018.

Copyright Notice

Copyright (c) 2018 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 (<https://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. Local Protection Overview	3
4. Extensions for the LSPA object	4
4.1. The Preference TLV	4
4.2. The Bypass TLV	5
4.3. The LOCALLY-PROTECTED-LSPS TLV	6
5. IANA considerations	8
5.1. PCEP-Error Object	8
5.2. PCEP TLV Type Indicators	8
6. Security Considerations	8
7. Contributors	8
8. References	9
8.1. Normative References	9
8.2. Informative References	9
Appendix A. Additional Stuff	10
Authors' Addresses	10

1. Introduction

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

Stateful PCE [RFC8231] 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 and allow delegation of control of LSPs to PCEs.

In a network where all LSPs have control delegated to a PCE, the PCE can apply global concurrent optimization to optimize LSP placement. The PCE can also control the timing and sequence of path computation and applying path changes. In a deployment where a PCE is used to compute all the paths, it may be beneficial for the protection paths to also be controlled through the PCE. This document defines extensions needed for the setup and management of protection paths by the PCE.

Benefits of stateful synchronization and control of the protection paths include:

- o Better control over traffic after a failure and more deterministic path computation of protection paths. The PCE can optimize the protection path based on data not available to the PCC, for instance the PCE can make sure the protection path will not violate the delay specified by [I-D.ietf-pce-pcep-service-aware].
- o Satisfy more complex constraints and diversity requirements, such as maintaining diverse paths for LSPs as well as their local protection paths.
- o Given the PCE's global view of network resources, act as a form of LSP admission control into a protection path to ensure links are not overloaded during failure events.
- o On a PLR with multiple available protection routes, allows the PCE to map LSPs to all available protection routes versus a single best protection route.
- o Most of the benefits stated in the stateful PCE applicability draft [I-D.ietf-pce-stateful-pce-app-04] apply equally to protection paths.

2. Terminology

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

This document uses the following terms defined in [RFC8231] Stateful PCE, Delegation, Delegation Timeout Interval, LSP State Report, LSP Update Request.

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

3. Architectural Overview

3.1. Local Protection Overview

Local protection refers to the ability to locally route around failure of an LSP. Two types of local protection are possible:

- (1) 1:1 protection - the protection path protects a single LSP.
- (2) N:1 protection - the protection path protects multiple LSPs traversing the protected resource.

It is assumed that the PCE knows what resources require protection through mechanisms outside the scope of this document. In a PCE controlled deployment, support of 1:1 protection has limited applicability, and can be achieved as a degenerate case of 1:N protection. For this reason, local protection will be discussed only for the N:1 case.

Local protection requires the setup of a bypass at the PLR. This bypass can be PCC-initiated and delegated, or PCE-initiated. In either case, the PLR MUST maintain a PCEP session to the PCE. A bypass identifier (the name of the bypass) is required for disambiguation as multiple bypasses are possible at the PLR. There two types Bypass LSPs mappings:

(1) Independent Bypass LSP Mapping: In this case Bypass LSP mapping is handled by a local policy on PCC and the PCC reports all mappings to the PCE. In other words, bypass LSP(s) are mapped to any protected LSP(s) that satisfy PCC local policy.

(2) Dependent Bypass LSP mapping: Mapping of LSPs to bypass is done through a new TLV, the LOCALLY-PROTECTED-LSPS TLV in the LSP Update message from PCE to PLR. See section Section 4.3. When an LSP requiring protection is set up through the PLR, the PLR checks if it has a mapping to a bypass and only provides protection if such a mapping exists. The status of bypasses and what LSPs are protected by them is communicated to the PCE via LSP Status Report messages.

4. Extensions for the LSPA object

4.1. The Preference TLV

When provisioning a PCC, the PCE can influence primary to bypass LSP association of the PCC using the preference TLV. Bypass LSPs with a higher preference are used first during primary LSP association. Bypass LSPs with identical preferences are used for primary association according to local PCC selection.

The format of the IPv4 Preference TLV is shown in the following figure:

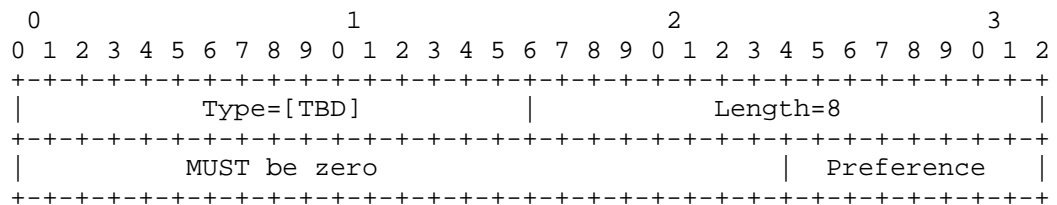


Figure 1: IPv4 Preference TLV format

The type of the TLV is [TBD] and it has a fixed length of 8 octets. The value contains the following fields:

Preference (8 bits): The value indicates the bypass LSP preference during the primary LSP selection process of the PCC. A lower preference value is preferred to a higher value with a default value of 255. A value of 0 would indicate that the bypass is not to be selected for any primary LSP associations.

If the Preference TLV is included, then the LSPA object MUST also carry the SYMBOLIC-PATH-NAME TLV as one of the optional TLVs. Failure to include the mandatory SYMBOLIC-PATH-NAME TLV MUST trigger PCErr of type 6 (Mandatory Object missing) and value TBD (SYMBOLIC-PATH-NAME TLV missing for bypass LSP).

4.2. The Bypass TLV

The facility backup method creates a bypass tunnel to protect a potential failure point. The bypass tunnel protects a set of LSPs with similar backup constraints [RFC4090].

A PCC can delegate a bypass tunnel to PCE control or a PCE can provision the bypass tunnel via a PCC. The procedures for bypass instantiation rely on the extensions defined in [RFC8281] and will be detailed in a future version of this document.

A subscription multiplier can be used to influence the local PCC admission control during primary LSP association. This allows for under subscription or oversubscription policy to be applied to the bandwidth attribute of the bypass LSP.

The Bypass TLV carries information about the bypass tunnel. It is included in the LSPA Object in LSP State Report and LSP Update Request messages.

The format of the IPv4 Bypass TLV is shown in the following figure:

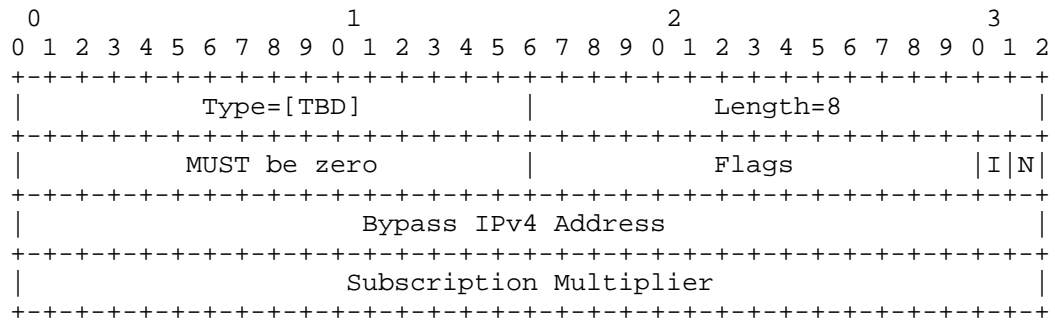


Figure 2: IPv4 Bypass TLV format

The type of the TLV is [TBD] and it has a fixed length of 8 octets. The value contains the following fields:

Flags (16 bit)

N (Node Protection - 1 bit): The N flag indicates whether the Bypass is used for node-protection. If the N flag is set to 1, the Bypass is used for node-protection. If the N flag is 0, the Bypass is used for link-protection.

I (Local Protection In Use - 1 bit): The I Flag indicates that local repair mechanism is in use.

Bypass IPv4 address: The Bypass IPv4 Address is the next-hop address of the protected link in the paths of the protected LSPs.

Subscription Multiplier (32 bits): An optional multiplier represented as a floating point number. The value may be used to influence CAC during primary LSP association. For example, a bypass may reserved 50M but the PCC may want to admit up to (multiplier * reserved bandwidth) to the bypass LSP.

If the Bypass TLV is included, then the LSPA object MUST also carry the SYMBOLIC-PATH-NAME TLV as one of the optional TLVs. Failure to include the mandatory SYMBOLIC-PATH-NAME TLV MUST trigger PCerr of type 6 (Mandatory Object missing) and value TBD (SYMBOLIC-PATH-NAME TLV missing for bypass LSP)

4.3. The LOCALLY-PROTECTED-LSPS TLV

The IPV4-LOCALLY-PROTECTED-LSPS TLV in the LSPA Object contains a list of LSPs protected by the bypass tunnel.

The format of the Locally protected LSPs TLV is shown in the following figure:

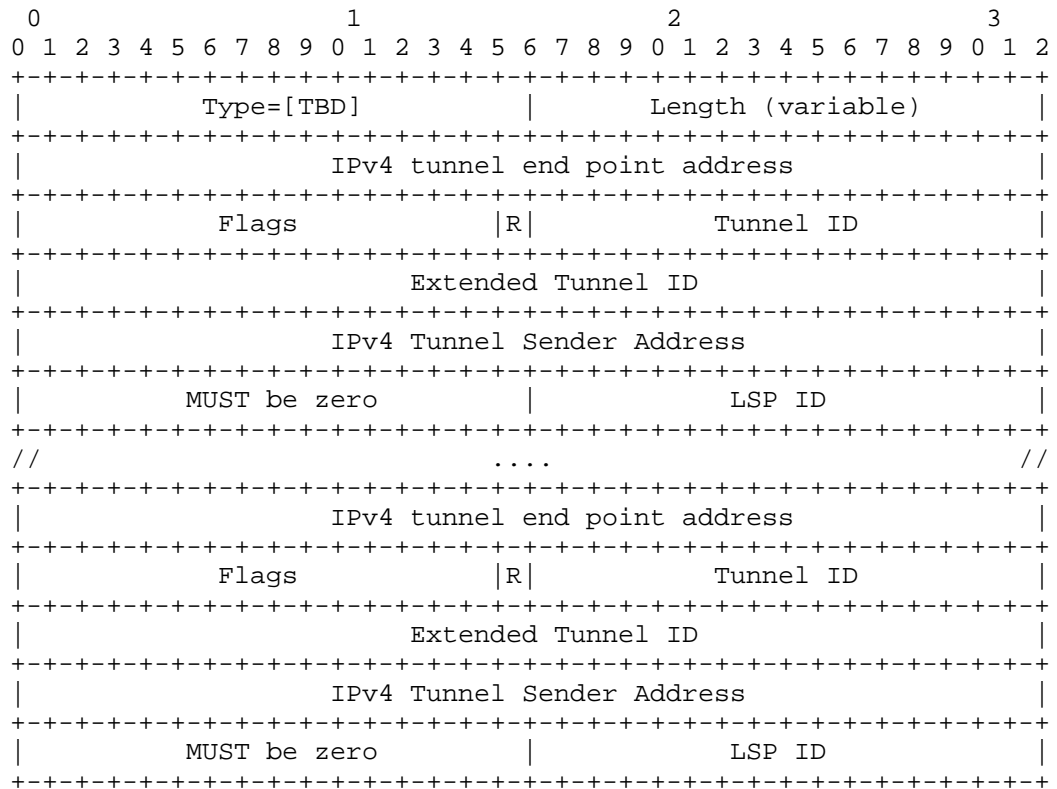


Figure 3: IPv4 Locally protected LSPs TLV format

The type of the TLV is [TBD] and it is of variable length. The value contains one or more LSP descriptors including the following fields filled per [RFC3209]

IPv4 Tunnel end point address: As defined in [RFC3209], Section 4.6.1.1

Flags (16 bit)

R(Remove - 1 bit): The R flag indicates that the LSP has been removed from the list of LSPs protected by the bypass tunnel.

Tunnel ID: As defined in [RFC3209], Section 4.6.1.1

Extended Tunnel ID: As defined in [RFC3209], Section 4.6.2.1

IPv4 Tunnel Sender address: As defined in [RFC3209], Section 4.6.2.1

LSP ID: As defined in RFC 3209

5. IANA considerations

5.1. PCEP-Error Object

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

Error-Type Meaning 6 Mandatory Object missing Error-value=TBD:
 SYMBOLIC-PATH-NAME TLV missing for a path where the S-bit is set in the LSPA object. Error-value=TBD: SYMBOLIC-PATH-NAME TLV missing for a bypass path.

5.2. PCEP TLV Type Indicators

This document defines the following new PCEP TLVs:

Value #	Meaning	Reference
???	Bypass	This Document
???	Weight	This Document
???	LOCALLY-PROTECTED-LSPS	This Document

Table 1: New PCEP TLVs

6. Security Considerations

The same security considerations apply at the PLR as those describe for the head end in PCE Initiated LSPs [RFC8281].

7. Contributors

The following people have substantially contributed to the editing of this document:

Harish Sitaraman, Juniper Networks, hsitaraman@juniper.net

Vishnu Pavan Beeram, Juniper Networks, vbeeram@juniper.net

Chandrasekar Ramachandran, Juniper Networks, csekar@juniper.net

Ambrose Kwong, Juniper Networks, akwong@juniper.net

Phil Bedard, bedard.phil@gmail.com

8. References

8.1. Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <<https://www.rfc-editor.org/info/rfc2119>>.
- [RFC2205] Braden, B., Zhang, L., Berson, S., Herzog, S., and S. Jamin, "Resource ReSerVation Protocol (RSVP) -- Version 1 Functional Specification", September 1997.
- [RFC3209] Awduche, D., Berger, L., Gan, D., Li, T., Srinivasan, V., and G. Swallow, "RSVP-TE: Extensions to RSVP for LSP Tunnels", December 2001.
- [RFC4090] Pan, P., Swallow, G., and A. Atlas, "Fast Reroute Extensions to RSVP-TE for LSP Tunnels", May 2005.
- [RFC5226] Narten, T. and H. Alvestrand, "Guidelines for Writing an IANA Considerations Section in RFCs", May 2008.
- [RFC5440] Vasseur, JP. and JL. Le Roux, "Path Computation Element (PCE) Communication Protocol (PCEP)", March 2009.
- [RFC8231] Crabbe, E., Medved, J., Minie, I., and R. Verga, "PCEP Extensions for Stateful PCE", 2015.
- [RFC8281] Crabbe, E., Sivabalan, S., and R. Verga, "PCEP Extensions for PCE-initiated LSP Setup in a Stateful PCE Model", 2014.

8.2. Informative References

- [I-D.narten-iana-considerations-rfc2434bis] Narten, T. and H. Alvestrand, "Guidelines for Writing an IANA Considerations Section in RFCs", draft-narten-iana-considerations-rfc2434bis-09 (work in progress), March 2008.
- [RFC2629] Rose, M., "Writing I-Ds and RFCs using XML", RFC 2629, DOI 10.17487/RFC2629, June 1999, <<https://www.rfc-editor.org/info/rfc2629>>.

- [RFC3552] Rescorla, E. and B. Korver, "Guidelines for Writing RFC Text on Security Considerations", BCP 72, RFC 3552, DOI 10.17487/RFC3552, July 2003, <<https://www.rfc-editor.org/info/rfc3552>>.
- [RFC4655] Farrel, A., Vasseur, J., and J. Ash, "A Path Computation Element (PCE)-Based Architecture", August 2006.
- [RFC4657] Ash, J. and J. Le Roux, "Path Computation Element (PCE) Communication Protocol Generic Requirements", September 2006.
- [RFC5394] Bryskin, I., Papadimitriou, D., Berger, L., and J. Ash, "Policy-Enabled Path Computation Framework", December 2008.
- [RFC5557] Lee, Y., Le Roux, J.L., King, D., and E. Oki, "Path Computation Element Communication Protocol (PCEP) Requirements and Protocol Extensions in Support of Global Concurrent Optimization", July 2009.

Appendix A. Additional Stuff

This becomes an Appendix.

Authors' Addresses

Colby Barth
Juniper Networks
Sunnyvale, CA
USA

Email: cbarth@juniper.net

Raveendra Torvi
Juniper Networks
Sunnyvale, CA
USA

Email: rtorvi@juniper.net

Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: March 7, 2016

X. Chen
Z. Li
Huawei Technologies
September 4, 2015

PCE-initiated IP Tunnel
draft-chen-pce-initiated-ip-tunnel-00

Abstract

This document specifies a set of extensions to PCEP to support PCE-initiated IP Tunnel to satisfy the requirement which is introduced in [I-D.li-spring-tunnel-segment]. The extensions include the setup, maintenance and teardown of PCE-initiated IP Tunnels, without the need for local configuration on the PCC.

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 March 7, 2016.

Copyright Notice

Copyright (c) 2015 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. Procedures for PCE-initiated IP Tunnel	3
3.1. Overview of Procedures	3
3.2. Capability Advertisement	4
3.3. Tunnel Operations	5
3.3.1. PCE IP Tunnel Instantiation	5
3.3.2. PCE IP Tunnel Update	5
3.3.3. PCE IP Tunnel Deletion	6
4. PCEP Messages	7
4.1. PCTunnelInitiate Message	7
4.2. PCTunnelUpd Message	8
4.3. PCTunnelRpt Message	9
5. PCEP Objects	10
5.1. OPEN Object	10
5.1.1. PCE Initiated Tunnel Capability TLV	10
5.2. SRP Object	11
5.3. TUNNEL Object	11
5.3.1. Tunnel Identifier TLV	12
5.3.2. Tunnel Name TLV	15
5.3.3. Tunnel Parameter TLV	16
5.3.4. Tunnel Attribute TLV	20
6. IANA Considerations	21
7. Security Considerations	21
8. References	21
8.1. Normative References	21
8.2. Informative References	22
Authors' Addresses	23

1. Introduction

[I-D.li-spring-tunnel-segment] introduces a new type of segment, Tunnel Segment, for the segment routing. Tunnel segment can be used to reduce SID stack depth of SR path, span the non-SR domain or provide differentiated services. The tunnel segment can be allocated for MPLS RSVP-TE tunnel, SR-TE tunnel or IP Tunnel.

[I-D.li-spring-tunnel-segment] introduces two ways to set up the tunnel which is used as tunnel segment: one is to configure tunnel on the device, the other is PCE-initiated tunnel.

[I-D.ietf-pce-stateful-pce], [I-D.ietf-pce-pce-initiated-lsp] and [I-D.ietf-pce-segment-routing] has defined how to set up the PCE initiated RSVP-TE LSP and SR-TE LSP. This document specifies a set of extensions to PCEP to support PCE-initiated IP Tunnel. The extensions include the setup, maintenance and teardown of PCE-initiated IP Tunnels, without the need for local configuration on the PCC.

2. Terminology

SR: Segment Routing

SR-TE: Segment Routing Traffic Engineering

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

The following terms are defined in [I-D.ietf-pce-pce-initiated-lsp]:

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

The following terms are defined in this document:

IP Tunnel: Tunnel that uses IP encapsulation.

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

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

3. Procedures for PCE-initiated IP Tunnel

3.1. Overview of Procedures

A PCC or PCE indicates its ability to support PCE Initiated dynamic tunnel during the PCEP Initialization Phase via "PCE Initiated Tunnel Capability" TLV (see details in Section 5.1).

In this document the procedure is only about PCE Initiated dynamic IP Tunnel. The decision when to instantiate or delete a PCE-initiated IP Tunnel is out of the scope of this document.

This section introduces the procedure to support PCE provisioned IP Tunnel as follows:

Firstly both the PCC and the PCE negotiate the PCE Initiated Tunnel Capability for tunnel types during the PCE session initiation phase. On the PCEP session with PCE Initiated Tunnel Capability PCE communicates with PCC to set up, maintain and tear down PCE-initiated IP Tunnels.

The procedure about tunnel state synchronization, PCC local policy and timeout process, the session failure process, etc. will be specified in the future version.

3.2. Capability Advertisement

During PCEP session establishment, both the PCC and the PCE must announce their support of PCEP extensions defined in this document. A PCEP Speaker (PCE or PCC) includes the "PCE Initiated Tunnel Capability" TLV, described in Section 5.1, in the OPEN Object to advertise its support for PCEP extensions for PCE Initiated IP Tunnel Capability.

The PCE Initiated Tunnel Capability TLV includes the tunnel types that are supported by PCEP Speaker. Each tunnel type is indicated by one bit.

The presence of the PCE Initiated Tunnel Capability TLV in PCE's OPEN message indicates that the PCE can support the instantiation of PCE-initiated Tunnels and the types of the tunnels which PCE can initiate.

The presence of such Capability TLV in PCC's OPEN Object indicates that the PCC can support to instantiate the tunnel according to the PCE's indication and the types of the tunnels which PCC can setup automatically according to the PCE's request.

If PCE has such capability TLV and PCC has no such capability TLV PCE MUST NOT send the PCE messages for procedure of PCE initiated IP Tunnel. And if PCC receives such messages it should send PCErr message to PCE.

If both PCE and PCC have such capability TLV they only negotiate the types of the tunnels both PCE and PCC can support. PCE MUST only initiate the specific tunnel which both PCE and PCC can support. Otherwise PCC MUST send the PCErr message.

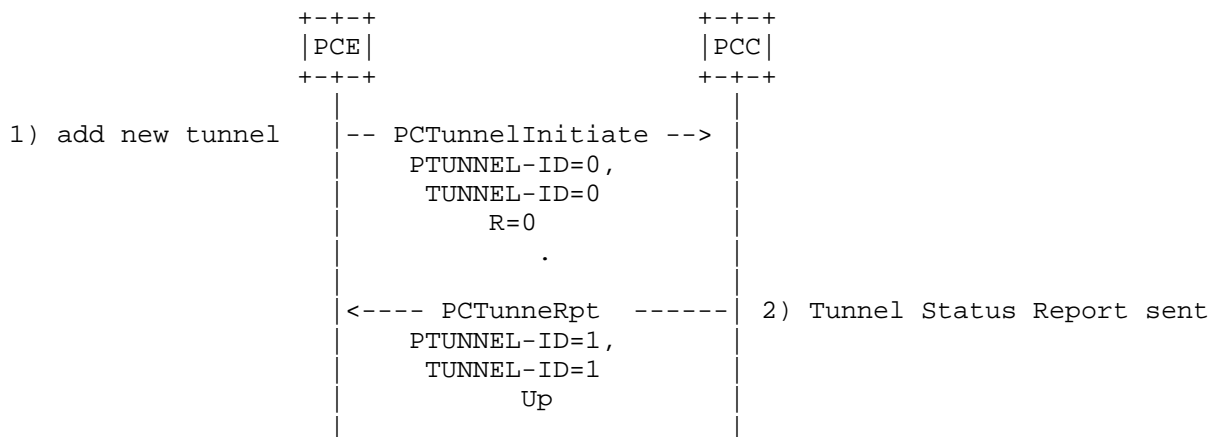
3.3. Tunnel Operations

3.3.1. PCE IP Tunnel Instantiation

To instantiate a tunnel, the PCE sends a Path Computation Tunnel Initiate (PCTunnelInitiate) message to the PCC. The PCTunnelInitiate message MUST include the SRP object (see details in Section 5.2) and TUNNEL object (see details in Section 5.3) . The TUNNEL object MUST have a PTUNNEL-ID of 0 and MUST include the Tunnel Identifier TLV with the TUNNEL-ID 0 and the Tunnel Name TLV. The TUNNEL object MAY have the Tunnel Parameter TLV.

The PCC creates the different type of tunnel using the end point address carried in Tunnel Identifier TLV and sends the Path Computation Tunnel State Report (PCTunneRpt) message to PCE. The PCTunneRpt message MUST include the SRP object and TUNNEL object. PCC assigns a unique PTUNNEL-ID carried via TUNNEL object and a unique TUNNEL-ID carried via Tunnel Identifier TLV(see details in Section 5.3) in TUNNEL object for the tunnel. PCC indicates the operational state in the TUNNEL object.

The PCTunneRpt message MUST include the SRP object, with the SRP-ID-NUMBER used in the SRP object of the PCTunnelInitiate message.

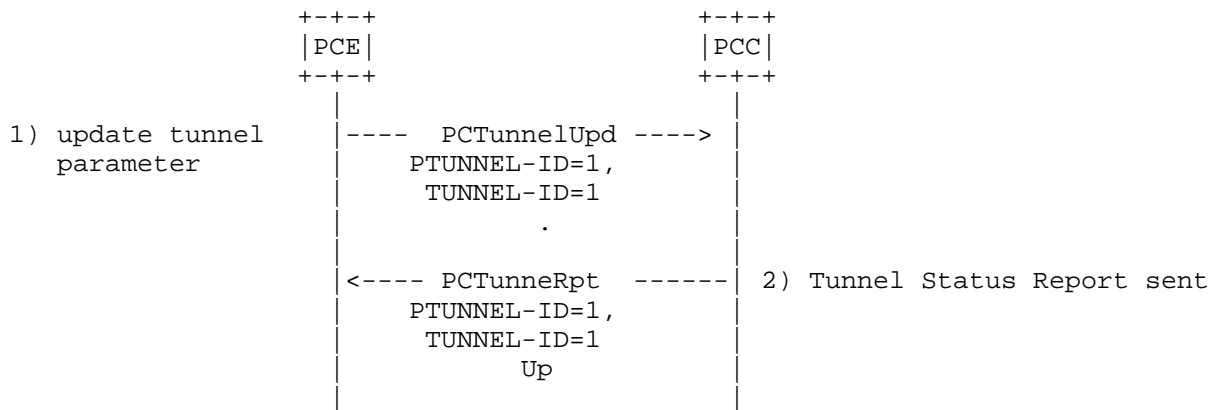


3.3.2. PCE IP Tunnel Update

To update the parameters used to create a tunnel, the PCE sends a Path Computation Tunnel Update (PCTunnelUpd) message to the PCC. The PCTunnelUpd message MUST include the SRP object and TUNNEL object. The TUNNEL object MUST have specific PTUNNEL-ID and MUST have specific Tunnel Identifier TLV. The TUNNEL object MUST carry any of the Tunnel Parameter TLV and Tunnel Attribute TLV.

The PCC updates the encapsulation parameters and/or attributes of the tunnel and PCC sends the PCTunneRpt message to PCE to report updated state.

The PCTunneRpt message MUST include the SRP object, with the SRP-ID-NUMBER used in the SRP object of the PCTunnelUpd message.

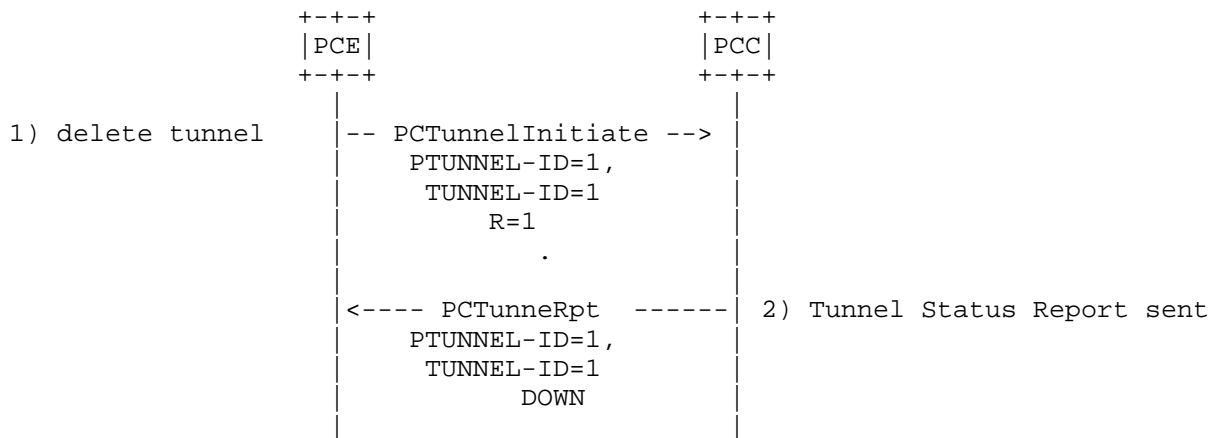


3.3.3. PCE IP Tunnel Deletion

To delete a tunnel, the PCE sends a Path Computation Tunnel Initiate (PCTunnelInitiate) message to the PCC. The PCTunnelInitiate message MUST include the SRP object and TUNNEL object and the 'R' flag in SRP object SHOULD be set. The TUNNEL object MUST have specific PTUNNEL-ID and MUST have specific Tunnel Identifier TLV.

The PCC delete the tunnel specified by PTUNNEL-ID and PCC sends the PCTunneRpt message to PCE to report updated state.

The PCTunneRpt message MUST include the SRP object, with the SRP-ID-NUMBER used in the SRP object of the PCTunnelInitiate message.



4. PCEP Messages

To initiate a tunnel, a PCE sends a PCTunnelInitiate message to a PCC.

To report the state of a tunnel, a PCC sends a PCTunnelRpt message to a PCE.

To modify the parameters of a tunnel, a PCE sends a PCTunnelUpd message to a PCC.

The message format, objects and TLVs are discussed separately below for the creation and the deletion cases.

4.1. PCTunnelInitiate Message

A Path Computation Tunnel Initiate message which is also referred to as PCTunnelInitiate message is a PCEP message sent by a PCE to a PCC to trigger tunnel instantiation or deletion.

The Message-Type field of the PCEP common header for the PCTunnelInitiate message is to be assigned by IANA. The PCTunnelInitiate message MUST include the SRP and the TUNNEL objects. If the SRP object is missing, the PCC MUST send a PCErr with error-type 6 (Mandatory Object missing) and error-value=10 (SRP Object missing) (per [I-D.ietf-pce-stateful-pce]). If the TUNNEL object is missing, the PCC MUST send a PCErr with error-type 6 (Mandatory Object missing) and error-value which means TUNNEL Object missing.

Tunnel instantiation is done by sending an Tunnel Initiate Message with an TUNNEL object with the reserved PTUNNEL-ID 0. Tunnel deletion is done by sending an Tunnel Initiate Message with an TUNNEL

object carrying the PTUNNEL-ID of the tunnel to be removed and an SRP object with the R flag set.

The format of a PCTunnelInitiate message for tunnel instantiation is as follows:

```
<PCTunnelInitiate Message> ::= <Common Header>
                                <PCE-initiated-tunnel-list>
```

Where:

```
<PCE-initiated-tunnel-list> ::= <PCE-initiated-tunnel-request>
                                [<PCE-initiated-tunnel-request>]
<PCE-initiated-tunnel-request> ::= (<PCE-initiated-tunnel-instantiation>
                                   |<PCE-initiated-tunnel-deletion>)
<PCE-initiated-tunnel-instantiation> ::= <SRP>
                                         <TUNNEL>
<PCE-initiated-tunnel-deletion> ::= <SRP>
                                     <TUNNEL>
```

The SRP object defined in [I-D.ietf-pce-stateful-pce] can be used in this document to correlate tunnel initiate requests and update requests sent by the PCE with the error reports and tunnel state reports sent by the PCC. Every request from the PCE sends a new SRP-ID-NUMBER. This number is unique per PCEP session and is incremented each time an operation (initiation, update, etc) is requested from the PCE. The value of the SRP-ID-NUMBER MUST be echoed back by the PCC in PCErr and PCTunnelRpt messages to allow for correlation between requests made by the PCE and errors or state reports generated by the PCC. Procedure of PCE-initiated IP Tunnel share the same number space of the SRP-ID-NUMBER with procedure of stateful PCE.

The <TUNNEL> object is an new object introduced in this document. <TUNNEL> object in PCTunnelInitiate message MUST include Tunnel Identifier TLV and Tunnel Name TLV. Tunnel Parameter TLV is optionally included.

The Tunnel Initiate message for tunnel instantiation has the TUNNEL object with the TUNNEL-ID in Tunnel Identifier TLV 0. The Tunnel Initiate message for tunnel deletion has the TUNNEL object carrying the TUNNEL-ID of the TUNNEL to be removed.

4.2. PCTunnelUpd Message

A Path Computation Tunnel Update Request message (also referred to as PCTunnelUpd message) is a PCEP message sent by a PCE to a PCC to update the encapsulation parameters and/or attributes of a tunnel. A PCTunnelUpd message can carry more than one Tunnel Update Request.

The Message-Type field of the PCEP common header for the PCUpd message is to be assigned by IANA.

The PCTunnelUpd message MUST include the SRP and the TUNNEL objects. If the SRP object is missing, the PCC MUST send a PCErr with error-type 6 (Mandatory Object missing) and error-value=10 (SRP Object missing) (per [I-D.ietf-pce-stateful-pce]). If the TUNNEL object is missing, the PCC MUST send a PCErr with error-type 6 (Mandatory Object missing) and error-value which means TUNNEL Object missing.

The format of a PCTunnelUpd message for tunnel parameter update is as follows:

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

Where:

```
<tunnel-update-request-list> ::= <tunnel-update-request>
                                [<tunnel-update-request-list>]
<tunnel-update-request> ::= <SRP>
                           <TUNNEL>
```

<TUNNEL> object in PCTunnelUpd message MUST include Tunnel Identifier TLV and any of Tunnel Parameter TLV and Tunnel Attribute TLV. Tunnel Name TLV is not included.

4.3. PCTunnelRpt Message

A Path Computation Tunnel State Report message which is also referred to as PCTunnelRpt message is a PCEP message sent by a PCC to a PCE to report the current state of a tunnel. A PCTunnelRpt message can carry more than one Tunnel State Reports. A PCC sends an Tunnel State Report in response to a Tunnel Initiate Request for creation or a Tunnel Update Request from a PCE.

The Message-Type field of the PCEP common header for the PCTunnelRpt message is to be assigned by IANA. The PCTunnelRpt message MUST include the SRP and the TUNNEL objects. If the SRP object is missing, the PCE MUST send a PCErr with error-type 6 (Mandatory Object missing) and error-value=10 (SRP Object missing) (per [I-D.ietf-pce-stateful-pce]). If the TUNNEL object is missing, the PCE MUST send a PCErr with error-type 6 (Mandatory Object missing) and error-value which means TUNNEL Object missing.

The format of a PCTunnelRpt message for tunnel instantiation is as follows:

```

<PCTunnelRpt Message> ::= <Common Header>
                           <tunnel-state-report-list>

```

Where:

```

<tunnel-state-report-list> ::= <tunnel-state-report>
                               [<tunnel-state-report-list>]
<tunnel-state-report> ::= <SRP>
                           <TUNNEL>

```

<TUNNEL> object in PCTunnelRpt message MUST include Tunnel Identifier TLV.

Tunnel Parameter TLV and Tunnel Attribute TLV is optionally included in PCTunnelRpt message. In the first PCTunnelRpt message in response to the PCTunnelInitiate message Tunnel Name TLV MUST be included. And in the subsequent PCTunnelRpt message Tunnel Name TLV is optionally included.

5. PCEP Objects

5.1. OPEN Object

5.1.1. PCE Initiated Tunnel Capability TLV

The PCE-INITIATE-TUNNEL-CAPABILITY TLV is an optional TLV associated with the OPEN Object [RFC5440] to exchange PCE-initiated tunnel capability of PCEP speakers.

Its format is shown in the following figure:

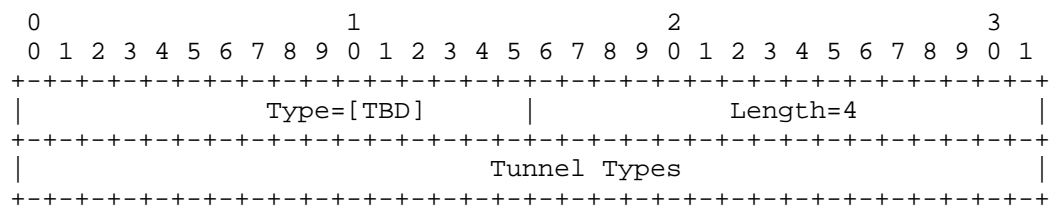


Figure 1: PCE-INITIATE-TUNNEL-CAPABILITY TLV

The type of the TLV is to be assigned by IANA and it has a fixed length of 4 octets.

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

Each bit indicates one kind of tunnel. Each bit from right to left successively represents the value of tunnel type which is 0 to 31. The value of tunnel types refer to the registry for "BGP Tunnel

Encapsulation Attribute Tunnel Types" [RFC5512] assigned by IANA.
This document only use the IP tunnel type.

The assignments used by this document are as follows:

Tunnel Type	Value
-----	-----
Reserved	0
GRE	2
VXLAN	8
NVGRE	9
MPLS in GRE	11
VxLAN GPE	12
MPLS in UDP	13

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

5.2. SRP Object

<SRP> object is defined in [I-D.ietf-pce-stateful-pce]. In this document <SRP> is used to correlate PCTunnelInitiate and PCTunnelRpt or PCErr message.

'R' Flag in <SRP> object is defined in [I-D.ietf-pce-pce-initiated-lsp]. When PCE requests PCC to create the IP tunnel 'R' Flag in <SRP> is set to 0. When PCE requests PCC to delete the IP tunnel 'R' Flag in <SRP> is set to 1.

Other flags must be set to 0 and if PCC receive the PCTunnelInitiate message with other reserved flags in <SRP > set to 1 PCC will send the PCErr message.

In procedure of PCE-initiated IP tunnel <SRP> object carries no optional TLVs.

5.3. TUNNEL Object

The TUNNEL object MUST be present within PCTunnelInitiate, PCTunnelRpt and PCTunnelUpd messages. The TUNNEL object contains a set of fields used to specify the target tunnel, the flags to indicate the state of the tunnel or operation to be performed on the tunnel and TLVs.

TUNNEL Object-Class is to be assigned by IANA.

TUNNEL Object-Type is 1.

The format of the TUNNEL object body is shown in following Figure:

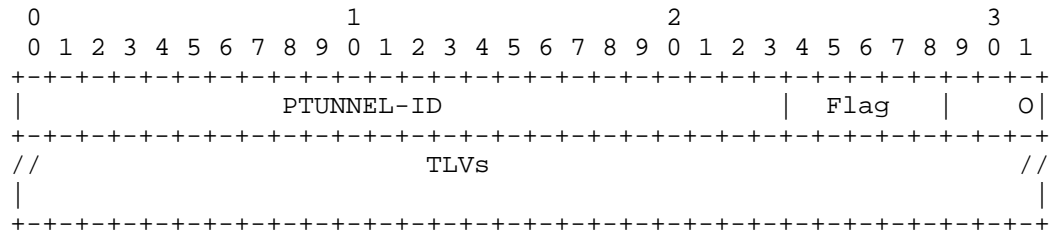


Figure 2: TUNNEL Object

PTUNNEL-ID (24 bits): A PCEP-specific identifier for the tunnel. A PCC creates a unique PTUNNEL-ID for each tunnel that is constant for the lifetime of a PCEP session. The PCC will advertise the same PTUNNEL-ID on all PCEP sessions. The mapping of the Tunnel Name to PTUNNEL-ID is communicated to the PCE by sending a PCTunnelRpt message containing the TUNNEL-NAME TLV. All subsequent PCEP messages then address the tunnel by the PTUNNEL-ID. The values of 0 and 0xFFFFFFFF are reserved.

Flags (8 bits):

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

The following values are defined:

0 - DOWN: The tunnel can't carry the traffic.

1 - UP: The tunnel can carry the traffic.

2-7 - 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.

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

5.3.1. Tunnel Identifier TLV

The Tunnel Identifier TLV MUST be included in the TUNNEL object in PCTunnelInitiate, PCTunnelRpt and PCTunnelUpd messages for PCE-initiated IP Tunnels. If the TLV is missing, the PCE will generate an error with error-type 6 (mandatory object missing) and error-value

which means Tunnel Identifier TLV missing and close the session. There are two Tunnel Identifier TLVs, one for IPv4 and one for IPv6.

The format of the IPV4-TUNNEL-Identifier TLV is shown in the following figure:

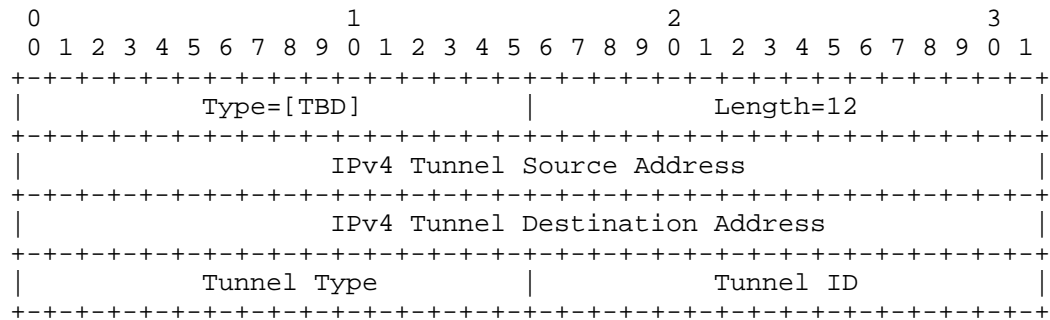


Figure 3: IPV4-TUNNEL-Identifier TLV

The type of the TLV is to be assigned by IANA and it has a fixed length of 12 octets. The value contains the following fields:

IPv4 Tunnel Source Address: contains the source IPv4 address of the ingress node of the tunnel.

IPv4 Tunnel Destination Address: contains the destination IPv4 address of the egress node of the tunnel.

Tunnel Type: contains the type of tunnel. The value of tunnel types refer to the registry for "BGP Tunnel Encapsulation Attribute Tunnel Types" [RFC5512] IANA set up. This document only use the IP tunnel type.

The assignments used by this document are as follows:

Tunnel Type	Value
-----	-----
Reserved	0
GRE	2
VXLAN	8
NVGRE	9
MPLS in GRE	11
VxLAN GPE	12
MPLS in UDP	13

Tunnel ID: Tunnel ID remains constant over the life time of a tunnel. A PCC creates a unique Tunnel ID for each tunnel. Each tunnel type

has individual identifier space. The Tunnel ID is allocated on id space of the tunnel type and is unique in the same id space.

The PCC will advertise the same Tunnel ID on all PCEP sessions. The mapping of the Tunnel Name to Tunnel ID is communicated to the PCE by sending a PCTunnelRpt message containing the TUNNEL-NAME TLV. The values of 0 and 0xFFFF are reserved.

The format of the IPV6-TUNNEL-Identifier TLV is shown in following figure:

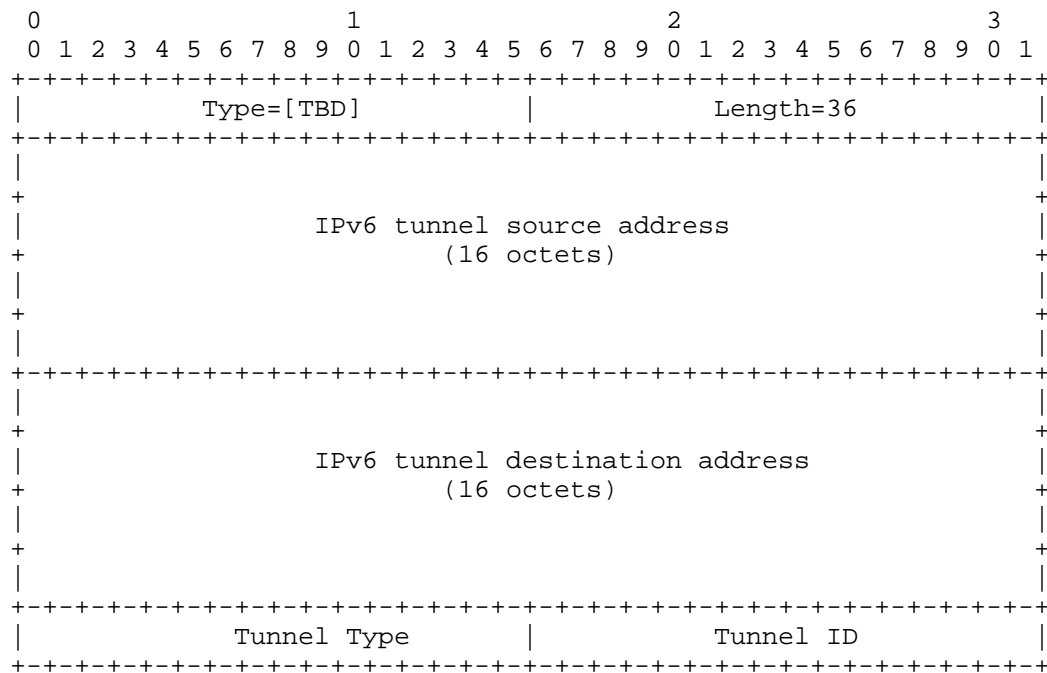


Figure 4: IPV6-TUNNEL-Identifier TLV

The type of the TLV is to be assigned by IANA and it has a fixed length of 36 octets. The value contains the following fields:

IPv6 Tunnel Source Address: contains the source IPv6 address of the ingress node of the tunnel.

IPv6 Tunnel Destination Address: contains the destination IPv6 address of the egress node of the tunnel.

Tunnel Type: contains the type of tunnel. The value of tunnel types refer to the registry for "BGP Tunnel Encapsulation Attribute Tunnel Types" [RFC5512] IANA set up.

This document only use the IP tunnel type. The assignments used by this document are as follows:

Tunnel Type	Value
-----	----
Reserved	0
GRE	2
VXLAN	8
NVGRE	9
MPLS in GRE	11
VxLAN GPE	12
MPLS in UDP	13

Tunnel ID: Tunnel ID remains constant over the life time of a tunnel. A PCC creates a unique Tunnel ID for each TUNNEL. Each tunnel type has individual identifier space. The tunnel ID is allocated on id space of the tunnel type and is unique in the same id space.

The PCC will advertise the same Tunnel ID on all PCEP sessions. The mapping of the Tunnel Name to Tunnel ID is communicated to the PCE by sending a PCTunnelRpt message containing the TUNNEL-NAME TLV. The values of 0 and 0xFFFF are reserved.

5.3.2. Tunnel Name TLV

The Tunnel Name TLV MUST be included in the TUNNEL object in PCTunnelInitiate messages for PCE-initiated IP Tunnels. If the TLV is missing, the PCE will generate an error with error-type 6 (mandatory object missing) and error-value which means Tunnel Name TLV missing and close the session.

Each tunnel MUST have a tunnel name that is unique in the PCC. This tunnel name MUST remain constant throughout a tunnel's lifetime.

The TUNNEL-NAME TLV MUST be included in the PCTunnelRpt message when a tunnel is first reported to a PCE in response to the PCTunnelInitiate message to create the tunnel. The tunnel name MAY be included in subsequent PCTunnelRpt messages for the tunnel.

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

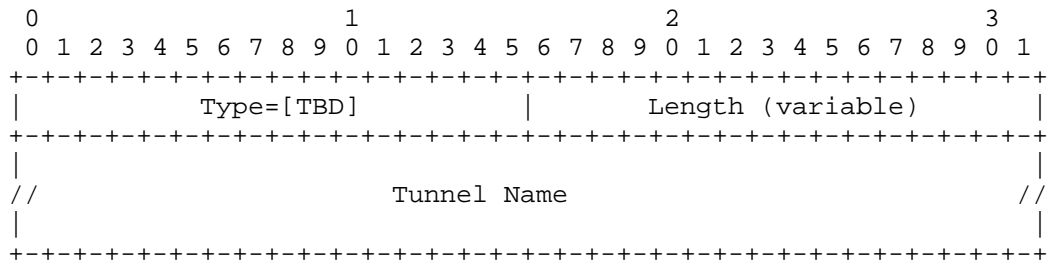


Figure 5: TUNNEL-NAME TLV

The type of the TLV is to be assigned by IANA and it has a variable length, which MUST be greater than 0.

5.3.3. Tunnel Parameter TLV

The Tunnel Parameter TLV and/or Tunnel Attribute TLV(see details in following section) MUST be included in the TUNNEL object in PCTunnelUpd messages for PCE-initiated IP Tunnels. If both of the TLVs are missing, the PCE will generate an error with error-type 6 (mandatory object missing) and error-value which means Tunnel Parameter TLV and Tunnel Attribute TLV missing and close the session.

The Tunnel Parameter TLV MAY be included in the TUNNEL object in PCTunnelInitiate and PCTunnelRpt messages for PCE-initiated IP Tunnels.

Tunnel Parameter TLV specifies information needed to construct the encapsulation header when sending packets through that tunnel.

The tunnel with different type has different encapsulation mode and each tunnel with same type MAY has different encapsulation parameters. When PCE initiate setup of the tunnel PCE can specify the encapsulation parameter of the tunnel and PCC will setup the tunnel and encapsulate the packet according to the parameters.

After the tunnel has been triggered to instantiate PCE can send PCTunnelUpd message to modify the encapsulation parameter.

The format of the TUNNEL-PARAMETER TLV is shown in following figure:

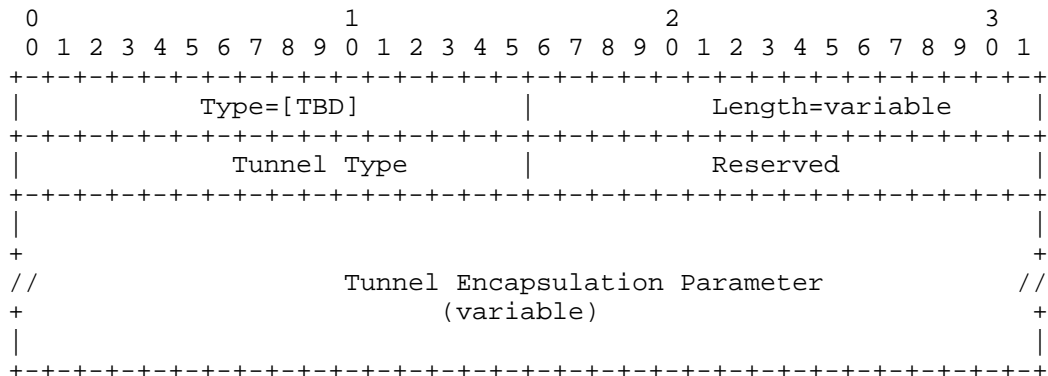


Figure 6: TUNNEL-PARAMETER TLV

The type of the TLV is to be assigned by IANA and it has a variable length, which MUST be greater than 0. The minimum value of length is 4 without any parameter. The value contains the following fields:

Tunnel Type: contains the type of tunnel. The value of tunnel types refer to the registry for "BGP Tunnel Encapsulation Attribute Tunnel Types" [RFC5512] IANA set up. This document only use the IP tunnel type.

The assignments used by this document are as follows:

Tunnel Type	Value
-----	-----
Reserved	0
GRE	2
VXLAN	8
NVGRE	9
MPLS in GRE	11
VxLAN GPE	12
MPLS in UDP	13

MPLS in GRE has the same encapsulation with GRE.

5.3.3.1. GRE

When the tunnel type of the TLV is GRE, the following is the structure of the value field of Tunnel Encapsulation Parameter:

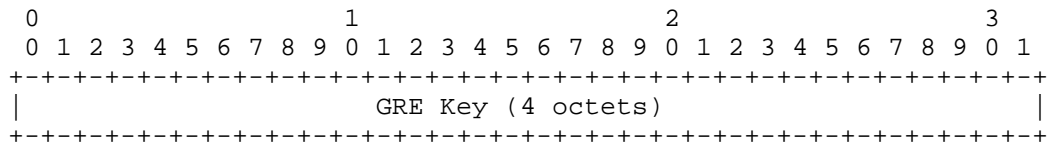


Figure 7: GRE Encapsulation TLV

* GRE Key: 4-octet field [RFC2890]. The actual method by which the key is obtained by PCE is beyond the scope of this document. The key is inserted into the GRE encapsulation header of the payload packets sent by ingress router to the egress router. It is intended to be used for identifying extra context information about the received payload.

Note that the key is optional. Unless a key value is being used, the GRE encapsulation MUST NOT be present. If GRE tunnel didn't use the GRE key the PCTunnelInitiate message needn't carry the TUNNEL-PARAMETER TLV. If GRE tunnel firstly use the GRE key the PCTunnelInitiate message need carry the TUNNEL-PARAMETER TLV. Then if the GRE tunnel quit using the GRE key the PCTunnelUpd message can carry the TUNNEL-PARAMETER TLV without GRE key to delete the parameter previously set.

5.3.3.2. VXLAN

When the tunnel type of the TLV is VXLAN, the following is the structure of the value field of Tunnel Encapsulation Parameter:

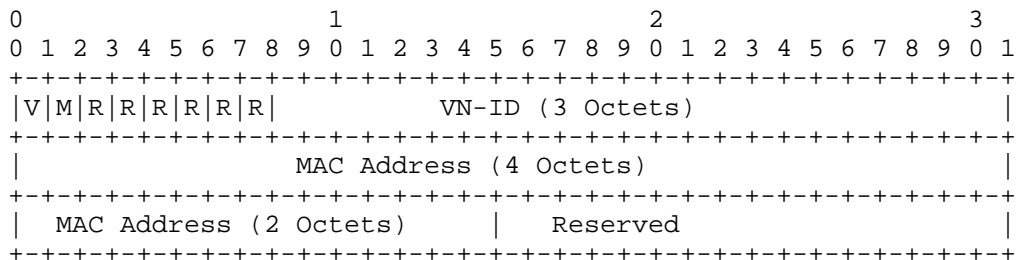


Figure 8: VXLAN Encapsulation TLV

The definition of the fields refer to [I-D.rosen-idr-tunnel-encaps].

5.3.3.3. VXLAN-GPE

When the tunnel type of the TLV is VXLAN-GPE, the following is the structure of the value field of Tunnel Encapsulation Parameter:

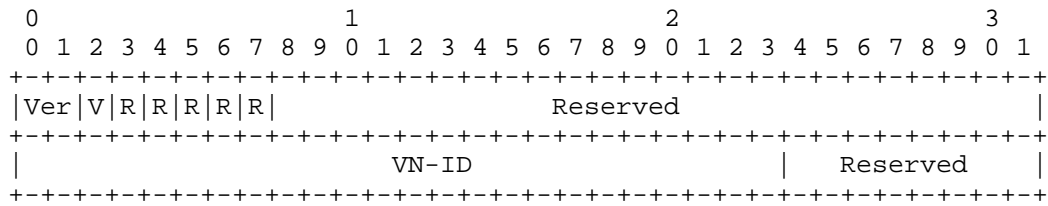


Figure 9: VXLAN GPE Encapsulation TLV

The definition of the fields refer to [I-D.rosen-idr-tunnel-encaps].

5.3.3.4. NVGRE

When the tunnel type of the TLV is NVGRE, the following is the structure of the value field of Tunnel Encapsulation Parameter:

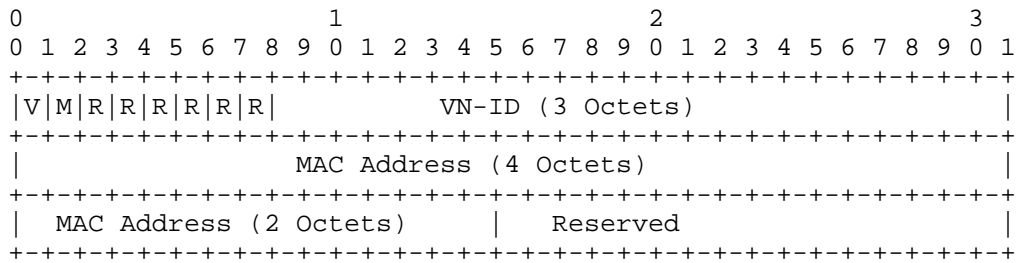


Figure 10: NVGRE Encapsulation TLV

The definition of the fields refer to [I-D.rosen-idr-tunnel-encaps].

5.3.3.5. MPLS-in-UDP

When the tunnel type of the TLV is MPLS-in-UDP, the following is the structure of the value field of Tunnel Encapsulation Parameter:

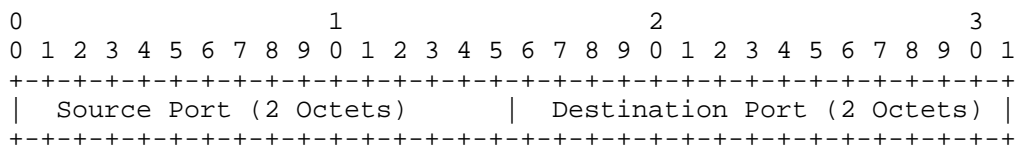


Figure 11: MPLS-in-UDP Encapsulation TLV

Source Port: UDP source port.

Destination Port: UDP destination port.

5.3.4. Tunnel Attribute TLV

The Tunnel Attribute TLV MAY be included in the TUNNEL object in PCTunnelInitiate, PCTunnelUpd, PCTunnelRpt messages for PCE-initiated IP Tunnels.

Tunnel Attribute TLV specifies some of the information of the tunnel such as metric or TE metric which are carried in sub-TLVs.

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

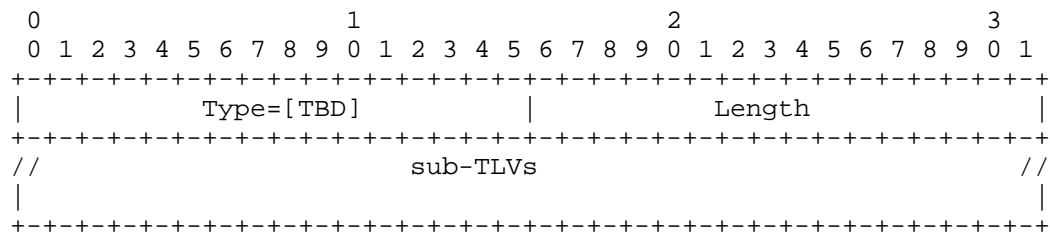


Figure 12: Tunnel Attribute TLV Format

The type of the TLV is to be assigned by IANA and it has a variable length. The minimum value of length is 0 without any parameter. The value contains the following fields:

sub-TLVs: Each sub-TLV has the Type(two octets), Length(two octets), Value. The length is the length of the value of the sub-TLV. Unknown sub-TLVs are to be ignored and skipped upon receipt.

This document defines the following sub-TLVs.

5.3.4.1. Metric Sub-TLV

The following is the structure of the sub-TLV of metric:

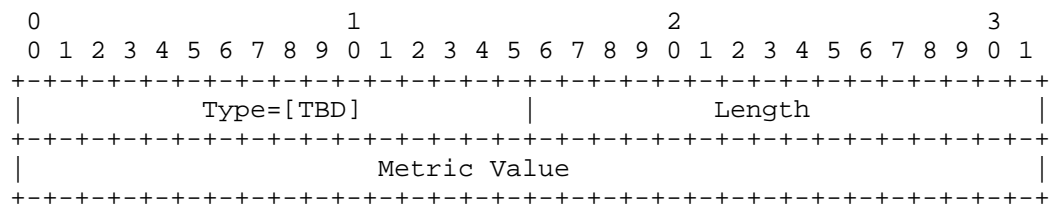


Figure 13: Metric Sub-TLV

The type of the sub-TLV is to be assigned by IANA and it has a fixed length of 4 octets.

The value comprises a single field - Metric Value (32 bits): value of metric.

5.3.4.2. TE Metric Sub-TLV

The following is the structure of the sub-TLV of traffic engineering metric:

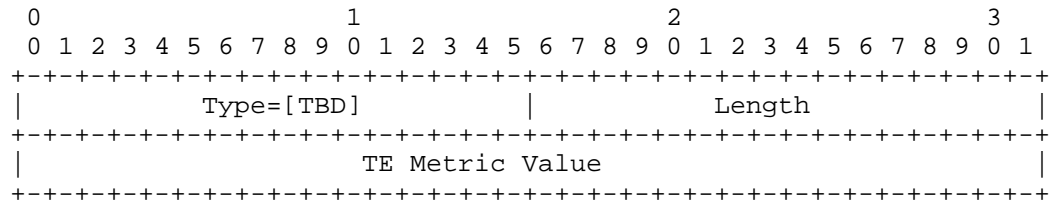


Figure 14: TE Metric Sub-TLV

The type of the sub-TLV is to be assigned by IANA and it has a fixed length of 4 octets.

The value comprises a single field - TE Metric Value (32 bits): value of traffic engineering metric.

6. IANA Considerations

TBD.

7. Security Considerations

TBD.

8. References

8.1. Normative References

[I-D.li-spring-tunnel-segment]

Li, Z. and N. Wu, "Tunnel Segment in Segment Routing", draft-li-spring-tunnel-segment-00 (work in progress), September 2015.

[I-D.rosen-idr-tunnel-encaps]

Rosen, E., Patel, K., and G. Velde, "Using the BGP Tunnel Encapsulation Attribute without the BGP Encapsulation SAFI", draft-rosen-idr-tunnel-encaps-03 (work in progress), August 2015.

- [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>>.
- [RFC2890] Dommety, G., "Key and Sequence Number Extensions to GRE", RFC 2890, DOI 10.17487/RFC2890, September 2000, <<http://www.rfc-editor.org/info/rfc2890>>.
- [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>>.
- [RFC5512] Mohapatra, P. and E. Rosen, "The BGP Encapsulation Subsequent Address Family Identifier (SAFI) and the BGP Tunnel Encapsulation Attribute", RFC 5512, DOI 10.17487/RFC5512, April 2009, <<http://www.rfc-editor.org/info/rfc5512>>.

8.2. Informative References

- [I-D.ietf-pce-pce-initiated-lsp]
Crabbe, E., Minei, I., Sivabalan, S., and R. Varga, "PCEP Extensions for PCE-initiated LSP Setup in a Stateful PCE Model", draft-ietf-pce-pce-initiated-lsp-04 (work in progress), April 2015.
- [I-D.ietf-pce-segment-routing]
Sivabalan, S., Medved, J., Filsfils, C., Crabbe, E., Lopez, V., Tantsura, J., Henderickx, W., and J. Hardwick, "PCEP Extensions for Segment Routing", draft-ietf-pce-segment-routing-06 (work in progress), August 2015.
- [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-11 (work in progress), April 2015.

Authors' Addresses

Xia Chen
Huawei Technologies
Huawei Bld., No.156 Beiqing Rd.
Beijing 100095
China

Email: jescia.chenxia@huawei.com

Zhenbin Li
Huawei Technologies
Huawei Bld., No.156 Beiqing Rd.
Beijing 100095
China

Email: lizhenbin@huawei.com

PCE Working Group
Internet-Draft
Intended status: Standards Track
Expires: December 18, 2015

D. Dhody
U. Palle
Huawei Technologies
R. Singh
Juniper Networks
R. Gandhi
Cisco Systems, Inc.
June 16, 2015

PCEP Extensions for MPLS-TE LSP Automatic Bandwidth Adjustment with
Stateful PCE
draft-dhody-pce-stateful-pce-auto-bandwidth-05

Abstract

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

This document describes automatic bandwidth adjustment of such LSPs when employing an Active Stateful PCE. In one of the models described, PCC computes the bandwidth to be adjusted and informs the PCE whereas in the second model, PCC reports the real-time traffic to a PCE and the PCE computes the adjustment bandwidth.

This document also describes automatic bandwidth adjustment for stateful PCE-initiated LSPs.

Status of This Memo

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

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

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

This Internet-Draft will expire on December 18, 2015.

Copyright Notice

Copyright (c) 2015 IETF Trust and the persons identified as the document authors. All rights reserved.

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

Table of Contents

1. Introduction	3
2. Conventions Used in This Document	5
2.1. Requirements Language	5
2.2. Terminology	5
3. Requirements for PCEP Extensions	5
4. Architectural Overview	7
4.1. Auto-Bandwidth Overview	7
4.2. Theory of Operation	9
4.3. Scaling Considerations	10
5. Extensions to the PCEP	10
5.1. AUTO-BANDWIDTH-ATTRIBUTE TLV	10
5.1.1. Adjustment Parameters	12
5.1.1.1. Sample-Interval sub-TLV	12
5.1.1.2. Adjustment-Interval sub-TLV	13
5.1.1.3. Adjustment Threshold	13
5.1.1.4. Minimum and Maximum Bandwidth	14
5.1.1.5. Overflow and Underflow Condition	15
5.1.2. Real-time Traffic Reporting	18
5.1.2.1. Real-time-Traffic-Report-Interval sub-TLV	19
5.1.2.2. Real-time-Traffic-Report-Threshold sub-TLV	19
5.1.2.3. Real-time-Traffic-Report-Threshold-Percentage sub-TLV	20
5.1.2.4. Real-time-Traffic-Report-Flow-Threshold sub-TLV	20
5.1.2.5. Real-time-Traffic-Report-Flow-Threshold- Percentage sub-TLV	21
5.2. BANDWIDTH Object	22
5.2.1. Auto-Bandwidth Adjusted Bandwidth	22
5.2.2. Bandwidth-Usage Report	22
5.3. The PCRpt Message	23

5.4. The PCInitiate Message	23
6. Security Considerations	23
7. Manageability Considerations	23
7.1. Control of Function and Policy	23
7.2. Information and Data Models	24
7.3. Liveness Detection and Monitoring	24
7.4. Verify Correct Operations	24
7.5. Requirements On Other Protocols	24
7.6. Impact On Network Operations	24
8. IANA Considerations	24
8.1. PCEP TLV Type Indicators	24
8.2. AUTO-BANDWIDTH-ATTRIBUTE Sub-TLV	24
8.3. BANDWIDTH Object	25
9. Acknowledgments	25
10. References	25
10.1. Normative References	25
10.2. Informative References	26
Appendix A. Contributor Addresses	27
Authors' Addresses	27

1. Introduction

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

[I-D.ietf-pce-stateful-pce] specifies extensions to PCEP to enable stateful control of MPLS TE LSPs. It describes two mode of operations - Passive Stateful PCE and Active Stateful PCE. In this document, the focus is on Active Stateful PCE where LSPs are configured at the PCC and control over them is delegated to the PCE. Further [I-D.ietf-pce-pce-initiated-lsp] describes the setup, maintenance and teardown of PCE-initiated LSPs under the stateful PCE model.

Over time, based on the varying traffic pattern, an LSP established with certain bandwidth may require to adjust the bandwidth, reserved in the network automatically. Ingress Label Switch Router (LSR) collects the traffic rate at each sample interval to determine the bandwidth demand of the LSP. This bandwidth information is then used to adjust the LSP bandwidth periodically. This feature is commonly referred to as Auto-Bandwidth.

Enabling Auto-Bandwidth feature on an LSP results in the LSP automatically adjusting its bandwidth based on the actual traffic flowing through the LSP. An LSP set-up with some arbitrary

(including zero) bandwidth value, automatically monitors the traffic flow and adjusts its bandwidth every adjustment-interval period. The bandwidth adjustment uses the make-before-break signaling method so that there is no interruption to traffic flow. This is described in detail in Section 4.1. [I-D.ietf-pce-stateful-pce-app] describes the use-case for Auto-Bandwidth adjustment for passive and active stateful PCE.

In this document, following deployment models are considered for employing Auto-Bandwidth feature with active stateful PCE.

o Deployment model 1: PCC to decide adjusted bandwidth:

- * In this model, the PCC (head-end of the LSP) monitors and calculates the new adjusted bandwidth. The PCC reports the calculated bandwidth to be adjusted to the PCE.
- * This approach would be similar to passive stateful PCE model, while the passive stateful PCE uses path request/reply mechanism, the active stateful PCE uses report/update mechanism to adjust the LSP bandwidth.
- * For PCE-initiated LSP, the PCC is requested during the LSP initiation to monitor and calculate the new adjusted bandwidth.

o Deployment model 2: PCE to decide adjusted bandwidth:

- * In this model, the PCE calculates the new adjusted bandwidth for the LSP.
- * Active stateful PCE can use information such as historical trending data, application-specific information about expected demands and central policy information along with real-time actual flow volumes to make smarter bandwidth adjustment to the delegated LSPs. Since the LSP has delegated control to the PCE, it is inherently suited that it should be the stateful PCE that decides the bandwidth adjustments.
- * For PCE-initiated LSP, the PCC is requested during initiation, to monitor and report the real-time bandwidth usage.
- * This model does not exclude use of any other mechanism employed by stateful PCE to learn real-time traffic information. But at the same time, using the same protocol (PCEP in this case) for updating and reporting the adjustment parameters as well as to learn real-time bandwidth usage is operationally beneficial.

This document defines extensions needed to support Auto-Bandwidth feature on the LSPs in a active stateful PCE model using PCEP.

2. Conventions Used in This Document

2.1. Requirements Language

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

2.2. Terminology

The following terminology is used in this document.

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

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

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

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

TE LSP: Traffic Engineering Label Switched Path.

Note the Auto-Bandwidth feature specific terms defined in Section 4.1.

3. Requirements for PCEP Extensions

There are two deployment models considered in this document for automatic bandwidth adjustments in case of active stateful PCE. In the model where PCC decides the adjusted bandwidth, PCC can report the new requested bandwidth and an active stateful PCE can update the bandwidth for a delegated LSP via existing mechanisms defined in [I-D.ietf-pce-stateful-pce]. Additional PCEP extensions required are summarized in the following table.

+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+

Model	PCC Initiated	PCE Initiated
PCC to decide adjusted bandwidth	<p>PCC monitors the traffic and reports the calculated bandwidth to be adjusted to the PCE.</p> <p>No new extensions are needed.</p> <p>Optionally AUTO-BANDWIDTH-ATTRIBUTE TLV can be used to identify the LSP with Auto-Bandwidth Feature enabled.</p>	<p>At the time of initiation, PCE request PCC to monitor the traffic and reports the calculated bandwidth to be adjusted to the PCE.</p> <p>Extension is needed for PCE to pass on the adjustment parameters at the time of Initiation.</p> <p>Refer the AUTO-BANDWIDTH-ATTRIBUTE TLV (and sub-TLVs e.g. Adjustment-Interval, Minimum-Bandwidth) in Section 5.1.</p>
PCC reports real-time traffic and PCE to decide adjusted bandwidth	<p>PCC monitors the traffic and reports the real-time traffic to the PCE. It is PCE that decides the calculated bandwidth to be adjusted and updates the LSP accordingly.</p> <p>Extension is needed for PCC to pass on the adjustment parameters at the time of delegation to PCE.</p> <p>Refer the AUTO-BANDWIDTH-ATTRIBUTE TLV (and sub-TLVs e.g. Adjustment-Threshold, Real-time-Traffic-Report-Interval) in Section 5.1.</p> <p>Further extension to</p>	<p>At the time of initiation, PCE request PCC to monitor the traffic and reports the real-time traffic to the PCE. It is PCE that decides the calculated bandwidth to be adjusted and updates the LSP accordingly.</p> <p>Extension is needed for PCE to pass on the real-time traffic reporting parameters at the time of Initiation.</p> <p>Refer the Real-time Traffic Reporting (e.g. Real-time-Traffic-Report-Interval, Real-time-Traffic-Report-Threshold) in Section 5.1.2.</p> <p>Further extension to report</p>

	report the real-time traffic to PCE are also needed (Refer Bandwidth-Usage type in Section 5.2.2).	the real-time traffic to PCE are also needed (Refer Bandwidth-Usage type in Section 5.2.2).
--	--	---

Table 1: Auto-Bandwidth Deployment Models

Additional Auto-Bandwidth deployment considerations are summarized below:

- o It is required to identify and inform the PCEP peer, the LSP that are enabled with Auto-Bandwidth feature. Not all LSPs in some deployments would like their bandwidth to be dependent on the real-time traffic but be constant as set by the operator.
- o It is also required to identify and inform the PCEP peer the model of operation i.e. if PCC decides the adjusted bandwidth, or PCC reports the real-time traffic instead and the PCE decides the adjusted bandwidth.
 - * Note that PCEP extension for reporting real-time traffic, as specified in this document, is one of the ways for a PCE to learn this information. But at the same time a stateful PCE may choose to learn this information from other means like management, performance tools, which are beyond the scope of this document.
- o Further for the LSP with Auto-Bandwidth feature enabled, an operator should be able to specify the adjustment parameters (i.e. configuration knobs) to control this feature (e.g. minimum/maximum bandwidth range) and PCEP peer should be informed.

4. Architectural Overview

4.1. Auto-Bandwidth Overview

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

The new bandwidth reservation is determined by sampling the actual traffic flowing through the LSP. If the traffic flowing through the LSP is lower than the configured or current bandwidth of the LSP, the

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

Note the following definitions of the Auto-Bandwidth terms:

Maximum Average Bandwidth (MaxAvgBw): The maximum average bandwidth represents the current traffic demand during a time interval. This is the maximum value of the averaged traffic rate in a given adjustment-interval.

Adjusted Bandwidth: This is the Auto-Bandwidth computed bandwidth that needs to be adjusted for the LSP.

Sample-Interval: The periodic time interval at which the traffic rate is collected as a sample.

Bandwidth-Sample (BwSample): The bandwidth sample collected at every sample interval to measure the traffic rate.

Adjustment-Interval: The periodic time interval at which the bandwidth adjustment should be made using the MaxAvgBw.

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

Minimum-Bandwidth: The minimum bandwidth that can be reserved for the LSP.

Adjustment-Threshold: This value is used to decide when the bandwidth should be adjusted. If the percentage or absolute difference between the current MaxAvgBw and the current bandwidth reservation is greater than or equal to the threshold value, the LSP bandwidth is adjusted to the current bandwidth demand (Adjusted Bandwidth) at the adjustment-interval expiry.

Overflow-Threshold: This value is used to decide when the bandwidth should be adjusted when there is a sudden increase in traffic demand. If the percentage or absolute difference between the current MaxAvgBw and the current bandwidth reservation is greater than or equal to the threshold value, the overflow-condition is set to be met. The LSP bandwidth is adjusted to the current

bandwidth demand bypassing the adjustment- interval if the overflow-condition is met consecutively for the overflow-counts.

Underflow-Threshold: This value is used to decide when the bandwidth should be adjusted when there is a sudden decrease in traffic demand. If the percentage or absolute difference between the current MaxAvgBw and the current bandwidth reservation is greater than or equal to the threshold value, the underflow-condition is set to be met. The LSP bandwidth is adjusted to the current bandwidth demand bypassing the adjustment- interval if the underflow-condition is met consecutively for the underflow-counts.

Report-Interval: This value indicates the periodic interval when the collected real-time traffic bandwidth samples (BwSample) should be reported to the stateful PCE via the PCRpt message.

Report-Threshold: This value is used to decide if the real-time traffic bandwidth samples collected should be reported. Only if the percentage or the absolute difference between at least one of the bandwidth samples collected and the current bandwidth reservation is greater than or equal to the threshold value, the bandwidth samples collected during the Report-Interval are reported otherwise the bandwidth sample(s) are skipped.

Report-Flow-Threshold: This value is used to decide when the real-time traffic bandwidth samples should be reported immediately when there is a sudden change in traffic demand. If the percentage or absolute difference between the current bandwidth sample and the current bandwidth reservation is greater than or equal to the flow threshold value, all the bandwidth samples collected so far are reported to the PCE immediately.

4.2. Theory of Operation

The traffic rate is periodically sampled at each sample-interval (which can be configured by the user and the default value as 5 minutes) by the head-end node of the LSP. The sampled traffic rates are accumulated over the adjustment-interval period (which can be configured by the user and the default value as 24 hours). The PCEP peer which is in-charge of calculating the bandwidth to be adjusted, will adjust the bandwidth of the LSP to the highest sampled traffic rate (MaxAvgBw) amongst the set of bandwidth samples collected over the adjustment-interval.

Note that the highest sampled traffic rate could be higher or lower than the current LSP bandwidth. Only if the difference between the current bandwidth demand (MaxAvgBw) and the current bandwidth reservation is greater than or equal to the Adjustment-Threshold

(percentage or absolute value), the LSP bandwidth is adjusted to the current bandwidth demand (MaxAvgBw).

In order to avoid frequent re-signaling, an operator may set a longer adjustment-interval value. However, longer adjustment-interval can result in an undesirable effect of masking sudden changes in traffic demands of an LSP. To avoid this, the Auto-Bandwidth feature may pre-maturely expire the adjustment-interval and adjust the LSP bandwidth to accommodate the sudden bursts of increase in traffic demand as an overflow condition or decrease in traffic demand as an underflow condition.

In case of Deployment model 2, the PCC reports the real-time traffic information and the PCE decides the adjusted bandwidth. Multiple bandwidth samples are collected every report-interval, and reported together to the PCE. To avoid reporting minor changes in real-time traffic, report-threshold is used, to suppress the sending of the collected samples during the report-interval. The collected samples are reported if at least one sample crosses the Report-Threshold (percentage or absolute value). In order to accommodate sudden changes in the real-time traffic, report flow threshold is employed by pre-maturely expiry of the report-interval to report the unreported bandwidth samples collected so far.

All thresholds in this document could be represented in both absolute value and percentage, and could be used together.

4.3. Scaling Considerations

There are potential scaling concerns for the model where PCC (ingress LSR) reports real-time traffic information to the stateful PCE for a large number of LSPs. It is recommended to combine multiple bandwidth samples (BwSample) using larger report-interval and report them together to the PCE, thus reducing the number of PCRpt messages. Further Report-Threshold can be use to skip reporting the bandwidth samples for small changes in the bandwidth.

The processing cost of monitoring a large number of LSPs at the PCC and handling bandwidth change requests at PCE should be taken into consideration. Note that, this will be implementation dependent.

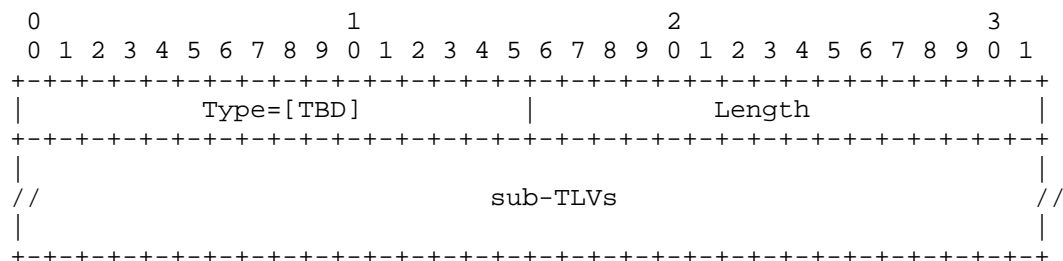
5. Extensions to the PCEP

5.1. AUTO-BANDWIDTH-ATTRIBUTE TLV

The AUTO-BANDWIDTH-ATTRIBUTE TLV can be included as an optional TLV in the LSPA object (as described in [RFC5440]). Whenever the LSP with Auto-Bandwidth feature enabled is delegated, AUTO-BANDWIDTH-

ATTRIBUTE TLV is carried in PCRpt message in LSPA object. The TLV provides PCE with the 'configurable knobs' of this feature. In case of PCE-Initiated LSP ([I-D.ietf-pce-pce-initiated-lsp]) with Auto-Bandwidth feature enabled, this TLV is included in LSPA object with PCInitiate message.

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



AUTO-BANDWIDTH-ATTRIBUTE TLV format

Type: TBD

Length: Variable

Value: This comprises one or more sub-TLVs.

Following sub-TLVs are defined in this document:

Type	Len	Name
1	4	Sample-Interval sub-TLV
2	4	Adjustment-Interval sub-TLV
3	4	Adjustment-Threshold sub-TLV
4	4	Adjustment-Threshold-Percentage sub-TLV
5	4	Minimum-Bandwidth sub-TLV
6	4	Maximum-Bandwidth sub-TLV
7	8	Overflow-Threshold sub-TLV
8	4	Overflow-Threshold-Percentage sub-TLV
9	8	Underflow-Threshold sub-TLV
10	4	Underflow-Threshold-Percentage sub-TLV
11	4	Real-time-Traffic-Report-Interval sub-TLV
12	4	Real-time-Traffic-Report-Threshold sub-TLV
13	4	Real-time-Traffic-Report-Threshold-Percentage sub-TLV
14	4	Real-time-Traffic-Report-Flow-Threshold sub-TLV
15	4	Real-time-Traffic-Report-Flow-Threshold-Percentage sub-TLV

Future specification can define additional sub-TLVs.

The presence of AUTO-BANDWIDTH-ATTRIBUTE TLV in LSPA object means that the automatic bandwidth adjustment feature is enabled. All sub-TLVs are optional and any unrecognized sub-TLV MUST be silently ignored. If a sub-TLV of same type appears more than once, only the first occurrence is processed and any others MUST be ignored.

If the sub-TLV are not encoded, the defaults based on the local policy are assumed.

The following sub-sections describe the sub-TLVs which are currently defined to be carried within the AUTO-BANDWIDTH-ATTRIBUTE TLV.

5.1.1.1. Adjustment Parameters

The sub-TLVs in this section are encoded to inform the PCEP peer the various sampling and adjustment parameters, and serves the following purpose -

- o For PCE-Initiated LSPs inform the PCC of the various sampling and adjustment parameters.
- o For PCC-Initiated LSPs in the Deployment Model 2 (where PCE decides the adjusted bandwidth), inform the PCE of the various sampling and adjustment parameters.

5.1.1.1.1. Sample-Interval sub-TLV

The Sample-Interval sub-TLV specifies a time interval in seconds at which traffic samples are collected at the PCC.

The Type is 1, Length is 4, and the value comprises of 4-octet time interval, the valid range is from 1 to 604800, in seconds. The default value is 300.

```

      0                   1                   2                   3
      0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-----+-----+-----+-----+-----+-----+-----+-----+
|                                     |                                     |
|          Type=1                     |          Length=4                |
|                                     |                                     |
|          Sample-Interval              |                               |
+-----+-----+-----+-----+-----+-----+-----+-----+

```

Sample-Interval sub-TLV format

5.1.1.2. Adjustment-Interval sub-TLV

The Adjustment-Interval sub-TLV specifies a time interval in seconds at which bandwidth adjustment should be made.

The Type is 2, Length is 4, and the value comprises of 4-octet time interval, the valid range is from 1 to 604800, in seconds. The default value is 300.

```

      0                   1                   2                   3
      0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-----+-----+-----+-----+-----+-----+-----+-----+
|                               Type=2                               |
+-----+-----+-----+-----+-----+-----+-----+-----+
|                               Adjustment-Interval                   |
+-----+-----+-----+-----+-----+-----+-----+-----+

```

Adjustment-Interval sub-TLV format

5.1.1.3. Adjustment Threshold

The sub-TLVs in this section are encoded to inform the PCEP peer the adjustment threshold parameters. An implementation MAY include both sub-TLVs for the absolute value and the percentage, in which case the bandwidth is adjusted when either of the adjustment threshold conditions are met.

5.1.1.3.1. Adjustment-Threshold sub-TLV

The Adjustment-Threshold sub-TLV is used to decide when the LSP bandwidth should be adjusted.

```

      0                   1                   2                   3
      0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-----+-----+-----+-----+-----+-----+-----+-----+
|                               Type=3                               |
+-----+-----+-----+-----+-----+-----+-----+-----+
|                               Adjustment Threshold                   |
+-----+-----+-----+-----+-----+-----+-----+-----+

```

Adjustment-Threshold sub-TLV format

The Type is 3, Length is 4, and the value comprises of -

- o Adjustment Threshold: The absolute Adjustment-Threshold bandwidth value, encoded in IEEE floating point format (see [IEEE.754.1985]), expressed in bytes per second. Refer to Section 3.1.2 of [RFC3471] for a table of commonly used values.

If the difference between the current MaxAvgBw and the current bandwidth reservation is greater than or equal to the threshold value, the LSP bandwidth is adjusted to the current bandwidth demand.

5.1.1.3.2. Adjustment-Threshold-Percentage sub-TLV

The Adjustment-Threshold-Percentage sub-TLV is used to decide when the LSP bandwidth should be adjusted.

```

      0               1               2               3
      0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-----+-----+-----+-----+-----+-----+-----+-----+
|               Type=4               |               Length=4               |
+-----+-----+-----+-----+-----+-----+-----+-----+
|               Reserved               |               Percentage               |
+-----+-----+-----+-----+-----+-----+-----+-----+

```

Adjustment-Threshold-Percentage sub-TLV format

The Type is 4, Length is 4, and the value comprises of -

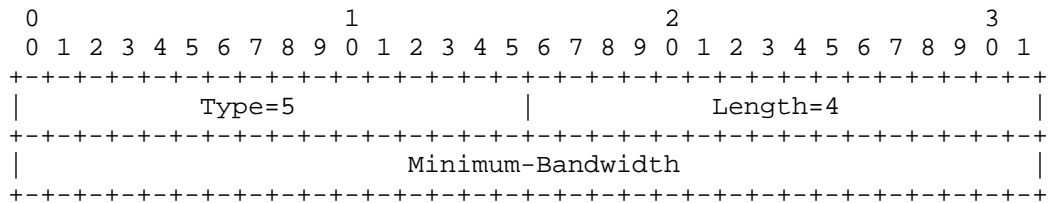
- o Reserved: SHOULD be set to zero on transmission and MUST be ignored on receipt.
- o Percentage: The Adjustment-Threshold value, encoded in percentage (an integer from 0 to 100). If the percentage difference between the current MaxAvgBw and the current bandwidth reservation is greater than or equal to the threshold percentage, the LSP bandwidth is adjusted to the current bandwidth demand.

5.1.1.4. Minimum and Maximum Bandwidth

5.1.1.4.1. Minimum-Bandwidth sub-TLV

The Minimum-Bandwidth sub-TLV specify the minimum bandwidth allowed for the LSP, and is expressed in bytes per second. The LSP bandwidth cannot be adjusted below the minimum bandwidth value.

The Type is 5, Length is 4, and the value comprises of 4-octet bandwidth value encoded in IEEE floating point format (see [IEEE.754.1985]), expressed in bytes per second. Refer to Section 3.1.2 of [RFC3471] for a table of commonly used values.

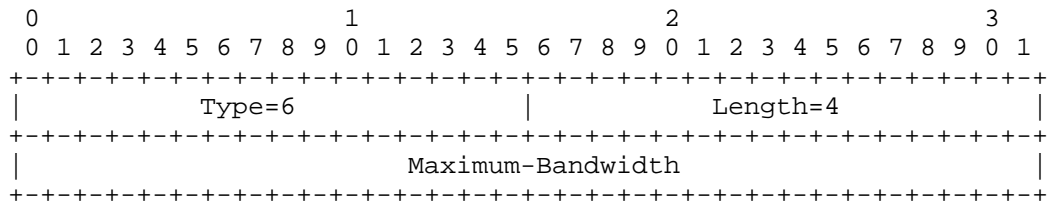


Minimum-Bandwidth sub-TLV format

5.1.1.4.2. Maximum-Bandwidth sub-TLV

The Maximum-Bandwidth sub-TLV specify the maximum bandwidth allowed for the LSP, and is expressed in bytes per second. The LSP bandwidth cannot be adjusted above the maximum bandwidth value.

The Type is 6, Length is 4, and the value comprises of 4-octet bandwidth value encoded in IEEE floating point format (see [IEEE.754.1985]), expressed in bytes per second. Refer to Section 3.1.2 of [RFC3471] for a table of commonly used values.



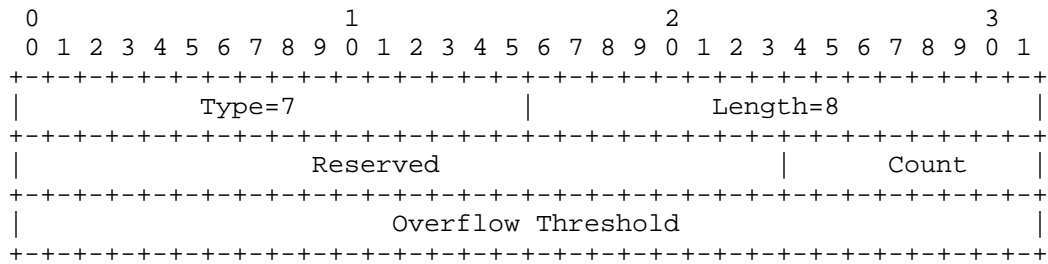
Maximum-Bandwidth sub-TLV format

5.1.1.5. Overflow and Underflow Condition

The sub-TLVs in this section are encoded to inform the PCEP peer the overflow and underflow threshold parameters. An implementation MAY include sub-TLVs for the absolute value and the percentage for the threshold, in which case the bandwidth is immediately adjusted when either of the adjustment threshold conditions are met consecutively for the given count.

5.1.1.5.1. Overflow-Threshold sub-TLV

The Overflow-Threshold sub-TLV is used to decide if the bandwidth should be adjusted immediately.



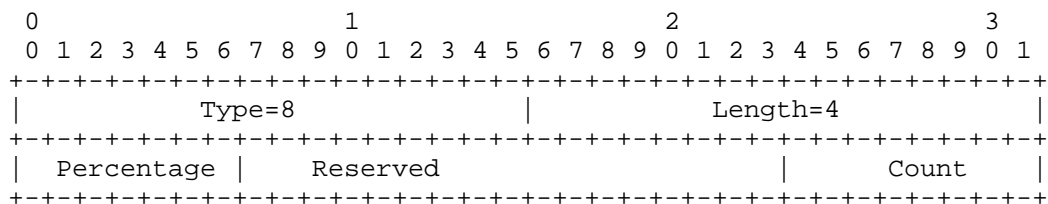
Overflow-Threshold sub-TLV format

The Type is 7, Length is 4, and the value comprises of -

- o Reserved: SHOULD be set to zero on transmission and MUST be ignored on receipt.
- o Count: The Overflow-Count value, encoded in integer. The value 0 is considered to be invalid. The number of consecutive samples for which the overflow condition MUST be met for the LSP bandwidth to be immediately adjusted to the current bandwidth demand, bypassing the adjustment-interval.
- o Overflow Threshold: The absolute Overflow-Threshold bandwidth value, encoded in IEEE floating point format (see [IEEE.754.1985]), expressed in bytes per second. Refer to Section 3.1.2 of [RFC3471] for a table of commonly used values. If the increase of the current MaxAvgBw from the current bandwidth reservation is greater than or equal to the threshold value, the overflow condition is met.

5.1.1.5.2. Overflow-Threshold-Percentage sub-TLV

The Overflow-Threshold-Percentage sub-TLV is used to decide if the bandwidth should be adjusted immediately.



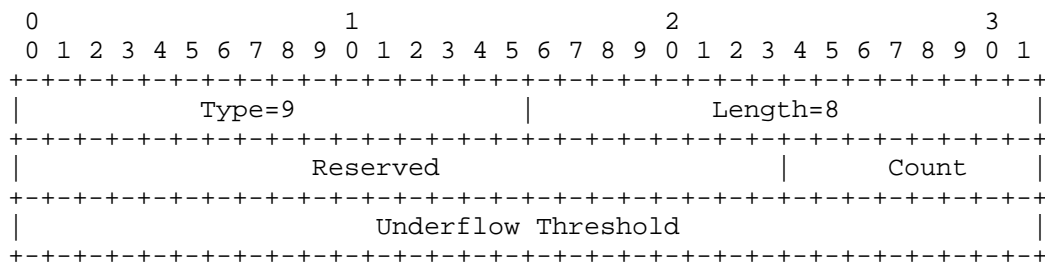
Overflow-Threshold-Percentage sub-TLV format

The Type is 8, Length is 4, and the value comprises of -

- o Percentage: The Overflow-Threshold value, encoded in percentage (an integer from 0 to 100). If the percentage increase of the current MaxAvgBw from the current bandwidth reservation is greater than or equal to the threshold percentage, the overflow condition is met.
- o Reserved: SHOULD be set to zero on transmission and MUST be ignored on receipt.
- o Count: The Overflow-Count value, encoded in integer. The value 0 is considered to be invalid. The number of consecutive samples for which the overflow condition MUST be met for the LSP bandwidth to be immediately adjusted to the current bandwidth demand, bypassing the adjustment-interval.

5.1.1.5.3. Underflow-Threshold sub-TLV

The Underflow-Threshold sub-TLV is used to decide if the bandwidth should be adjusted immediately.



Underflow-Threshold sub-TLV format

The Type is 9, Length is 8, and the value comprises of -

- o Reserved: SHOULD be set to zero on transmission and MUST be ignored on receipt.
- o Count: The Underflow-Count value, encoded in integer. The value 0 is considered to be invalid. The number of consecutive samples for which the underflow condition MUST be met for the LSP bandwidth to be immediately adjusted to the current bandwidth demand, bypassing the adjustment-interval.
- o Underflow Threshold: The absolute Underflow-Threshold bandwidth value, encoded in IEEE floating point format (see [IEEE.754.1985]), expressed in bytes per second. Refer to Section 3.1.2 of [RFC3471] for a table of commonly used values. If the decrease of the current MaxAvgBw from the current bandwidth

reservation is greater than or equal to the threshold value, the underflow condition is met.

5.1.1.5.4. Underflow-Threshold-Percentage sub-TLV

The Underflow-Threshold-Percentage sub-TLV is used to decide if the bandwidth should be adjusted immediately.

```

      0               1               2               3
      0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-----+-----+-----+-----+-----+-----+-----+-----+
|                                     |                                     |
|          Type=10                   |          Length=4                 |
+-----+-----+-----+-----+-----+-----+-----+-----+
| Percentage |      Reserved      |                                     |
|                                     |          Count          |
+-----+-----+-----+-----+-----+-----+-----+-----+

```

Underflow-Threshold-Percentage sub-TLV format

The Type is 10, Length is 4, and the value comprises of -

- o Percentage: The Underflow-Threshold value, encoded in percentage (an integer from 0 to 100). If the percentage decrease of the current MaxAvgBw from the current bandwidth reservation is greater than or equal to the threshold percentage, the underflow condition is met.
- o Reserved: SHOULD be set to zero on transmission and MUST be ignored on receipt.
- o Count: The Underflow-Count value, encoded in integer. The value 0 is considered to be invalid. The number of consecutive samples for which the underflow condition MUST be met for the LSP bandwidth to be immediately adjusted to the current bandwidth demand, bypassing the adjustment-interval.

5.1.2. Real-time Traffic Reporting

The sub-TLVs in this section are encoded to inform the PCEP peer the various real-time traffic reporting parameters in the Deployment Model 2 (where PCE decides the adjusted bandwidth). In this model, Real-time-Traffic-Report-Interval sub-TLV MUST be included to specify the frequency of reporting.

The report threshold is used to decide if the collected bandwidth samples should be reported or skipped. An implementation MAY include both sub-TLVs for the absolute value and the percentage, in which case the real-time traffic is reported when either of the report threshold conditions are met.

The report flow threshold is used to decide when the collected bandwidth samples should be reported immediately, bypassing the report interval. An implementation MAY include both sub-TLVs for the absolute value and the percentage, in which case the real-time traffic is reported immediately when either of the report flow threshold conditions are met.

5.1.2.1. Real-time-Traffic-Report-Interval sub-TLV

The Real-time-Traffic-Report-Interval sub-TLV specifies a time interval in seconds in which collected bandwidth samples should be reported to PCE.

The Type is 11, Length is 4, and the value comprises of 4-octet time interval, the valid range is from 1 to 604800, in seconds.

```

      0                   1                   2                   3
      0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-----+-----+-----+-----+-----+-----+-----+-----+
|                                     |                                     |
|          Type=11                   |          Length=4                 |
+-----+-----+-----+-----+-----+-----+-----+-----+
|                                     |                                     |
|          Real-time-Traffic-Report-Interval                           |
+-----+-----+-----+-----+-----+-----+-----+-----+

```

Real-time-Traffic-Report-Interval sub-TLV format

There is no default value. This sub-TLV MUST be included to enable the real-time traffic reporting.

5.1.2.2. Real-time-Traffic-Report-Threshold sub-TLV

The Real-time-Traffic-Report-Threshold sub-TLV is used to decide when the bandwidth samples collected should be reported immediately, bypassing the report-interval.

```

      0                   1                   2                   3
      0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-----+-----+-----+-----+-----+-----+-----+-----+
|                                     |                                     |
|          Type=12                   |          Length=4                 |
+-----+-----+-----+-----+-----+-----+-----+-----+
|                                     |                                     |
|          Real-time-Traffic-Report Threshold                           |
+-----+-----+-----+-----+-----+-----+-----+-----+

```

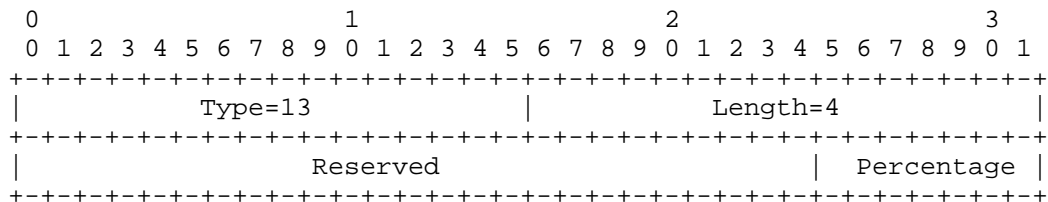
Real-time-Traffic-Report-Threshold sub-TLV format

The Type is 12, Length is 4, and the value comprises of -

- o **Threshold:** The absolute threshold bandwidth value, encoded in IEEE floating point format (see [IEEE.754.1985]), expressed in bytes per second. Refer to Section 3.1.2 of [RFC3471] for a table of commonly used values. If the increase or the decrease of at least one of the bandwidth samples (BwSample) collected so far compared to the current bandwidth reservation is greater than or equal to the threshold value, the bandwidth samples collected so far are reported.

5.1.2.3. Real-time-Traffic-Report-Threshold-Percentage sub-TLV

The Real-time-Traffic-Report-Threshold sub-TLV is used to decide when the bandwidth samples collected should be reported immediately, bypassing the report-interval.



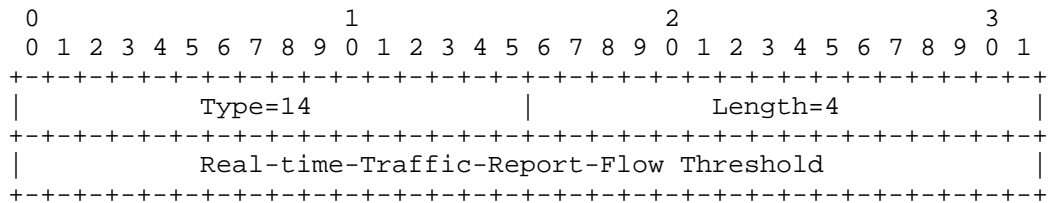
Real-time-Traffic-Report-Threshold-Percentage sub-TLV format

The Type is 13, Length is 4, and the value comprises of -

- o **Reserved:** SHOULD be set to zero on transmission and MUST be ignored on receipt.
- o **Percentage:** The threshold value, encoded in percentage (an integer from 0 to 100). If the percentage increase or the decrease of at least one of the bandwidth sample (BwSample) compared to the current bandwidth reservation is greater than or equal to the threshold percentage, the bandwidth samples collected so far are reported.

5.1.2.4. Real-time-Traffic-Report-Flow-Threshold sub-TLV

The Real-time-Traffic-Report-Flow-Threshold sub-TLV is used to decide when the bandwidth samples collected should be reported immediately, bypassing the report-interval.



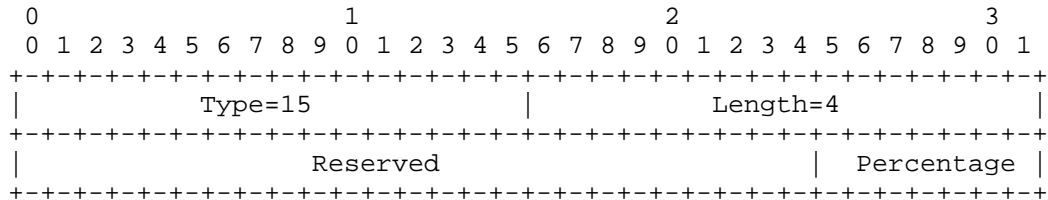
Real-time-Traffic-Report-Flow-Threshold sub-TLV format

The Type is 14, Length is 4, and the value comprises of -

- o Threshold: The absolute flow threshold bandwidth value, encoded in IEEE floating point format (see [IEEE.754.1985]), expressed in bytes per second. Refer to Section 3.1.2 of [RFC3471] for a table of commonly used values. If the increase or the decrease of the current bandwidth sample (BwSample) compared to the current bandwidth reservation is greater than or equal to the flow threshold value, all the bandwidth samples collected so far are reported immediately, bypassing the report-interval.

5.1.2.5. Real-time-Traffic-Report-Flow-Threshold-Percentage sub-TLV

The Real-time-Traffic-Report-Flow-Threshold sub-TLV is used to decide when the bandwidth samples collected should be reported immediately, bypassing the report-interval.



Real-time-Traffic-Report-Flow-Threshold-Percentage sub-TLV format

The Type is 15, Length is 4, and the value comprises of -

- o Reserved: SHOULD be set to zero on transmission and MUST be ignored on receipt.
- o Percentage: The flow threshold value, encoded in percentage (an integer from 0 to 100). If the percentage increase or the decrease of the current bandwidth sample (BwSample) compared to the current bandwidth reservation is greater than or equal to the threshold percentage, all the bandwidth samples collected so far are reported immediately, bypassing the report-interval.

5.2. BANDWIDTH Object

5.2.1. Auto-Bandwidth Adjusted Bandwidth

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

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

In the first model, where PCC calculates the adjusted bandwidth, PCC only reports the calculated bandwidth to be adjusted (MaxAvgBw) to the PCE. This is done via the existing 'Requested Bandwidth with BANDWIDTH Object-Type as 1'.

5.2.2. Bandwidth-Usage Report

A new BANDWIDTH object type is defined to report the actual bandwidth usage of a TE LSP.

The Object type is [TBD], the object body has a variable length, multiples of 4 bytes. The payload format is as follows:

```

      0               1               2               3
      0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-----+-----+-----+-----+-----+-----+-----+-----+
|                                     BwSample1                             |
+-----+-----+-----+-----+-----+-----+-----+-----+
|                                     ...                                     |
+-----+-----+-----+-----+-----+-----+-----+-----+
|                                     BwSampleN                             |
+-----+-----+-----+-----+-----+-----+-----+-----+

```

Bandwidth-Usage format

- o BwSample: The actual bandwidth usage, (the BwSample collected at the end of each sample-interval) encoded in IEEE floating point format (see [IEEE.754.1985]), expressed in bytes per second.

The Bandwidth-Usage object can be used in the second deployment model where PCC reports the TE LSP bandwidth usage and the PCE decides the auto-bandwidth adjusted bandwidth.

The Bandwidth-Usage object can also be used for TE LSPs without enabling the auto-bandwidth feature, to learn the actual bandwidth

usage of the LSPs for other applications at the stateful PCE. The details of which are beyond the scope of this document.

5.3. The PCRpt Message

When LSP is delegated to a PCE for the very first time, BANDWIDTH object of type 1 is used to specify the requested bandwidth in the PCRpt message.

When the LSP is enabled with the Auto-Bandwidth feature, and Real-time-Traffic-Report-Interval sub-TLV is not present (Deployment model 1), PCC SHOULD include the BANDWIDTH object of type 1 to specify the calculated bandwidth to be adjusted to the PCE in the PCRpt message.

When the LSP is enabled with the Auto-Bandwidth feature, and Real-time-Traffic-Report-Interval sub-TLV is present (Deployment model 2), PCC SHOULD include the BANDWIDTH object of type [TBD] to report the real-time traffic to the PCE in the PCRpt message.

The definition of the PCRpt message (see [I-D.ietf-pce-stateful-pce]) is unchanged by this document.

5.4. The PCInitiate Message

For PCE-initiated LSP [I-D.ietf-pce-pce-initiated-lsp] with Auto-Bandwidth feature enabled, AUTO-BANDWIDTH-ATTRIBUTE TLV MUST be included in LSPA object with the PCInitiate message. The rest of the processing remains unchanged.

6. Security Considerations

This document defines a new BANDWIDTH type and AUTO-BANDWIDTH-ATTRIBUTE TLV which do not add any new security concerns beyond those discussed in [RFC5440] and [I-D.ietf-pce-stateful-pce] in itself.

Some deployments may find the reporting of the real-time traffic information as extra sensitive and thus should employ suitable PCEP security mechanisms like TCP-AO or [I-D.ietf-pce-pceps].

7. Manageability Considerations

7.1. Control of Function and Policy

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

7.2. Information and Data Models

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

7.3. Liveness Detection and Monitoring

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

7.4. Verify Correct Operations

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

7.5. Requirements On Other Protocols

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

7.6. Impact On Network Operations

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

8. IANA Considerations

8.1. PCEP TLV Type Indicators

This document defines the following new PCEP TLVs; IANA is requested to make the following allocations from this registry.
<http://www.iana.org/assignments/pcep/pcep.xhtml#pcep-tlv-type-indicators>

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

8.2. AUTO-BANDWIDTH-ATTRIBUTE Sub-TLV

This document specifies the AUTO-BANDWIDTH-ATTRIBUTE Sub-TLVs. IANA is requested to create an "AUTO-BANDWIDTH-ATTRIBUTE Sub-TLV Types" sub-registry in the "PCEP TLV Type Indicators" for the sub-TLVs carried in the AUTO-BANDWIDTH-ATTRIBUTE TLV. This document defines the following types:

Type	Name	Reference
0	Reserved	[This I.D.]
1	Sample-Interval sub-TLV	[This I.D.]
2	Adjustment-Interval sub-TLV	[This I.D.]
3	Adjustment-Threshold sub-TLV	[This I.D.]
4	Adjustment-Threshold-Percentage sub-TLV	[This I.D.]
5	Minimum-Bandwidth sub-TLV	[This I.D.]
6	Maximum-Bandwidth sub-TLV	[This I.D.]
7	Overflow-Threshold sub-TLV	[This I.D.]
8	Overflow-Threshold-Percentage sub-TLV	[This I.D.]
9	Underflow-Threshold sub-TLV	[This I.D.]
10	Underflow-Threshold-Percentage sub-TLV	[This I.D.]
11	Real-time-Traffic-Report-Interval sub-TLV	[This I.D.]
12	Real-time-Traffic-Report-Threshold sub-TLV	[This I.D.]
13	Real-time-Traffic-Report-Threshold-Percentage sub-TLV	[This I.D.]
14	Real-time-Traffic-Report-Flow-Threshold sub-TLV	[This I.D.]
15	Real-time-Traffic-Report-Flow-Threshold-Percentage sub-TLV	[This I.D.]
16-65535	Unassigned	[This I.D.]

8.3. BANDWIDTH Object

This document defines new object type for the BANDWIDTH object; IANA is requested to make the following allocations from this registry.
<http://www.iana.org/assignments/pcep/pcep.xhtml#pcep-objects>

Object-Class Value	Name	Reference
5	BANDWIDTH	[This I.D.]
	Object-Type	
	TBD: Bandwidth-Usage Report	

9. Acknowledgments

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

10. References

10.1. Normative References

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

- [RFC5440] Vasseur, JP. and JL. Le Roux, "Path Computation Element (PCE) Communication Protocol (PCEP)", RFC 5440, March 2009.
- [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-11 (work in progress), April 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-04 (work in progress), April 2015.
- [IEEE.754.1985]
Institute of Electrical and Electronics Engineers,
"Standard for Binary Floating-Point Arithmetic", IEEE
Standard 754, August 1985.

10.2. Informative References

- [RFC3471] Berger, L., "Generalized Multi-Protocol Label Switching (GMPLS) Signaling Functional Description", RFC 3471, January 2003.
- [RFC7420] Koushik, A., Stephan, E., Zhao, Q., King, D., and J. Hardwick, "Path Computation Element Communication Protocol (PCEP) Management Information Base (MIB) Module", RFC 7420, December 2014.
- [I-D.ietf-pce-stateful-pce-app]
Zhang, X. and I. Minei, "Applicability of a Stateful Path Computation Element (PCE)", draft-ietf-pce-stateful-pce-app-04 (work in progress), April 2015.
- [I-D.ietf-pce-pceps]
Lopez, D., Dios, O., Wu, W., and D. Dhody, "Secure Transport for PCEP", draft-ietf-pce-pceps-04 (work in progress), May 2015.

Appendix A. Contributor Addresses

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

Email: kinghe@tencent.com

Xian Zhang
Huawei Technologies
Research Area F3-1B,
Huawei Industrial Base,
Shenzhen, 518129, China

Phone: +86-755-28972645
Email: zhang.xian@huawei.com

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

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

Authors' Addresses

Dhruv Dhody
Huawei Technologies
Divyashree Techno Park, Whitefield
Bangalore, Karnataka 560037
India

EMail: dhruv.ietf@gmail.com

Udayasree Palle
Huawei Technologies
Divyashree Techno Park, Whitefield
Bangalore, Karnataka 560037
India

EMail: udayasree.palle@huawei.com

Ravi Singh
Juniper Networks
1194 N. Mathilda Ave.
Sunnyvale, CA 94089
USA

EMail: ravis@juniper.net

Rakesh Gandhi
Cisco Systems, Inc.

EMail: rgandhi@cisco.com

PCE Working Group
Internet-Draft
Intended status: Standards Track
Expires: May 20, 2017

D. Dhody
U. Palle
Huawei Technologies
R. Singh
Juniper Networks
R. Gandhi
Individual Contributor
L. Fang
eBay
November 16, 2016

PCEP Extensions for MPLS-TE LSP Automatic Bandwidth Adjustment with
Stateful PCE
draft-dhody-pce-stateful-pce-auto-bandwidth-09

Abstract

The Path Computation Element Communication Protocol (PCEP) provides mechanisms for Path Computation Elements (PCEs) to perform path computations in response to Path Computation Clients (PCCs) requests. The stateful PCE extensions allow stateful control of Multi-Protocol Label Switching (MPLS) Traffic Engineering Label Switched Paths (TE LSPs) using PCEP.

This document describes PCEP extensions for automatic bandwidth adjustment when employing an Active Stateful PCE for both PCE-initiated and PCC-initiated 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."

Copyright Notice

Copyright (c) 2016 IETF Trust and the persons identified as the

document authors. All rights reserved.

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

Table of Contents

1. Introduction	3
2. Conventions Used in This Document	4
2.1. Requirements Language	4
2.2. Terminology	4
3. Requirements for PCEP Extensions	5
4. Architectural Overview	6
4.1. Auto-Bandwidth Overview	6
4.2. Auto-bandwidth Theory of Operation	8
4.3. Scaling Considerations	8
5. Extensions to the PCEP	9
5.1. Capability Advertisement	9
5.1.1 AUTO-BANDWIDTH-CAPABILITY TLV	9
5.2. AUTO-BANDWIDTH-ATTRIBUTE TLV	10
5.2.1. Sample-Interval sub-TLV	11
5.2.2. Adjustment-Interval sub-TLV	12
5.2.3. Adjustment Threshold	12
5.2.3.1. Adjustment-Threshold sub-TLV	12
5.2.3.2. Adjustment-Threshold-Percentage sub-TLV	13
5.2.4. Minimum and Maximum Bandwidth Values	13
5.2.4.1. Minimum-Bandwidth sub-TLV	13
5.2.4.2. Maximum-Bandwidth sub-TLV	14
5.2.5. Overflow and Underflow Conditions	14
5.2.5.1. Overflow-Threshold sub-TLV	14
5.2.5.2. Overflow-Threshold-Percentage sub-TLV	15
5.2.5.3. Underflow-Threshold sub-TLV	16
5.2.5.4. Underflow-Threshold-Percentage sub-TLV	17
5.3. BANDWIDTH Object	17
5.4. The PCInitiate Message	18
5.5. The PCRpt Message	18
5.6. The PCNtf Message	18
6. Security Considerations	19

7.	Manageability Considerations	19
7.1.	Control of Function and Policy	19
7.2.	Information and Data Models	19
7.3.	Liveness Detection and Monitoring	19
7.4.	Verify Correct Operations	20
7.5.	Requirements On Other Protocols	20
7.6.	Impact On Network Operations	20
8.	IANA Considerations	21
8.1.	PCEP TLV Type Indicators	21
8.2.	AUTO-BANDWIDTH-CAPABILITY TLV Flag Field	21
8.3.	AUTO-BANDWIDTH-ATTRIBUTE Sub-TLV	21
8.4.	Error Object	22
8.5.	Notification Object	22
9.	References	22
9.1.	Normative References	22
9.2.	Informative References	23
	Acknowledgments	24
	Contributors' Addresses	24
	Authors' Addresses	25

1. Introduction

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

[I-D.ietf-pce-stateful-pce] specifies extensions to PCEP to enable stateful control of MPLS TE LSPs. It describes two mode of operations - Passive stateful PCE and Active stateful PCE. In this document, the focus is on Active stateful PCE where LSPs are configured at the PCC and control over them is delegated to the PCE. Further [I-D.ietf-pce-pce-initiated-lsp] describes the setup, maintenance and teardown of PCE-initiated LSPs for the stateful PCE model.

Over time, based on the varying traffic pattern, an LSP established with certain bandwidth may require to adjust the bandwidth, reserved in the network automatically. Ingress Label Switch Router (LSR) collects the traffic rate at each sample interval to determine the bandwidth demand of the LSP. This bandwidth information is then used to adjust the LSP bandwidth periodically. This feature is commonly referred to as Auto-Bandwidth.

Enabling Auto-Bandwidth feature on an LSP results in the LSP automatically adjusting its bandwidth reservation based on the actual traffic flowing through the LSP. The initial LSP bandwidth can be set to an arbitrary value (including zero), in practice, it can be operator expected value based on design and planning. Once the LSP is set-up, the LSP monitors the traffic flow and adjusts its bandwidth every adjustment-interval period. The bandwidth adjustment uses the make-before-break signaling method so that there is no interruption to the traffic flow. The Auto-Bandwidth is described in detail in Section 4.1. [I-D.ietf-pce-stateful-pce-app] describes the use-case for Auto-Bandwidth adjustment for passive and active stateful PCE.

- o The PCC (head-end of the LSP) monitors and calculates the new adjusted bandwidth. The PCC reports the calculated bandwidth to be adjusted to the PCE.
- o This approach would be similar to passive stateful PCE model, while the passive stateful PCE uses path request/reply mechanism, the active stateful PCE uses report/update mechanism to adjust the LSP bandwidth.
- o For PCE-initiated LSP, the PCC is requested during the LSP initiation to monitor and calculate the new adjusted bandwidth.

This document defines extensions needed to support Auto-Bandwidth feature on the LSPs in a active stateful PCE model using PCEP.

2. Conventions Used in This Document

2.1. Requirements Language

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

2.2. Terminology

The following terminology is used in this document.

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

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

are delegated from a PCC to a PCE.

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

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

TE LSP: Traffic Engineering Label Switched Path.

Note the Auto-Bandwidth feature specific terms defined in Section 4.1.

3. Requirements for PCEP Extensions

The PCEP speaker supporting this document MUST have a mechanism to advertise the automatic bandwidth adjustment capability.

PCEP extensions required are summarized in the following table.

PCC Initiated	PCE Initiated
PCC monitors the traffic and reports the calculated bandwidth to be adjusted to the PCE.	At the time of initiation, PCE request PCC to monitor the traffic and report the calculated bandwidth to be adjusted to the PCE.
No new extensions are needed.	Extension is needed for PCE to pass on the adjustment parameters at the time of Initiation.

Table 1: Auto-Bandwidth PCEP extensions

Further Auto-Bandwidth deployment considerations are summarized below:

- o It is required to identify and inform the PCEP peer, the LSP that are enabled with Auto-Bandwidth feature. Not all LSPs in some

deployments would like their bandwidth to be dependent on the real-time bandwidth usage but be constant as set by the operator.

- o Further for the LSP with Auto-Bandwidth feature enabled, an operator should be able to specify the adjustment parameters (i.e. configuration knobs) to control this feature (e.g. minimum/maximum bandwidth range) and PCEP peer should be informed.

4. Architectural Overview

4.1. Auto-Bandwidth Overview

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

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

Note the following definitions of the Auto-Bandwidth terms:

Maximum Average Bandwidth (MaxAvgBw): The maximum average bandwidth represents the current traffic bandwidth demand during a time interval. This is the maximum value of the averaged traffic bandwidth rate in a given adjustment-interval.

Adjusted Bandwidth: This is the Auto-Bandwidth computed bandwidth that needs to be adjusted for the LSP.

Sample-Interval: The periodic time interval at which the traffic rate is collected as a sample.

Bandwidth-Sample (BwSample): The bandwidth sample collected at every sample interval to measure the traffic rate.

Adjustment-Interval: The periodic time interval at which the

bandwidth adjustment should be made using the MaxAvgBw.

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

Minimum-Bandwidth: The minimum bandwidth that can be reserved for the LSP.

Adjustment-Threshold: This value is used to decide when the bandwidth should be adjusted. If the percentage or absolute difference between the current MaxAvgBw and the current bandwidth reservation is greater than or equal to the threshold value, the LSP bandwidth is adjusted to the current bandwidth demand (Adjusted Bandwidth) at the adjustment-interval expiry.

Overflow-Count: This value is used to decide when the bandwidth should be adjusted when there is a sudden increase in traffic demand. This value indicates how many times consecutively, the percentage or absolute difference between the current MaxAvgBw and the current bandwidth reservation is greater than or equal to the Overflow-Threshold value.

Overflow-Threshold: This value is used to decide when the bandwidth should be adjusted when there is a sudden increase in traffic demand. If the percentage or absolute difference between the current MaxAvgBw and the current bandwidth reservation is greater than or equal to the threshold value, the overflow-condition is set to be met. The LSP bandwidth is adjusted to the current bandwidth demand bypassing the adjustment-interval if the overflow-condition is met consecutively for the Overflow-Count.

Underflow-Count: This value is used to decide when the bandwidth should be adjusted when there is a sudden decrease in traffic demand. This value indicates how many times consecutively, the percentage or absolute difference between the current MaxAvgBw and the current bandwidth reservation is greater than or equal to the Underflow-Threshold value.

Underflow-Threshold: This value is used to decide when the bandwidth should be adjusted when there is a sudden decrease in traffic demand. If the percentage or absolute difference between the current MaxAvgBw and the current bandwidth reservation is greater than or equal to the threshold value, the underflow-condition is set to be met. The LSP bandwidth is adjusted to the current bandwidth demand bypassing the adjustment-interval if the underflow-condition is met consecutively for the Underflow-Count.

4.2. Auto-bandwidth Theory of Operation

The traffic rate is periodically sampled at each sample-interval (which can be configured by the user and the default value as 5 minutes) by the head-end node of the LSP. The sampled traffic rates are accumulated over the adjustment-interval period (which can be configured by the user and the default value as 24 hours). The PCEP peer which is in-charge of calculating the bandwidth to be adjusted, will adjust the bandwidth of the LSP to the highest sampled traffic rate (MaxAvgBw) amongst the set of bandwidth samples collected over the adjustment-interval.

Note that the highest sampled traffic rate could be higher or lower than the current LSP bandwidth. Only if the difference between the current bandwidth demand (MaxAvgBw) and the current bandwidth reservation is greater than or equal to the Adjustment-Threshold (percentage or absolute value), the LSP bandwidth is adjusted to the current bandwidth demand (MaxAvgBw). Some LSPs are less eventful while other LSPs may encounter a lot of changes in the traffic pattern. PCE sets the intervals for adjustment based on the traffic pattern of the LSP.

In order to avoid frequent re-signaling, an operator may set a longer adjustment-interval value. However, longer adjustment-interval can result in an undesirable effect of masking sudden changes in traffic demands of an LSP. To avoid this, the Auto-Bandwidth feature may pre-maturely expire the adjustment-interval and adjust the LSP bandwidth to accommodate the sudden bursts of increase in traffic demand as an overflow condition or decrease in traffic demand as an underflow condition.

All thresholds in this document could be represented in both absolute value and percentage, and could be used together.

4.3. Scaling Considerations

It should be noted that any bandwidth change would require re-signaling of an LSP in a make-before-break fashion, which can further trigger preemption of lower priority LSPs in the network. When deployed under scale, this can lead to a signaling churn in the network. The Auto-bandwidth application algorithm is thus advised to take this into consideration before adjusting the LSP bandwidth. Operators are advised to set the values of various auto-bandwidth adjustment parameters appropriate for the deployed LSP scale.

If a PCE gets overwhelmed, it can notify the PCC to temporarily suspend its auto-bandwidth reporting (see Section 5.6). Similarly if

a PCC gets overwhelmed due to signaling churn, it can notify the PCE to temporarily suspend the LSP bandwidth adjustment.

5. Extensions to the PCEP

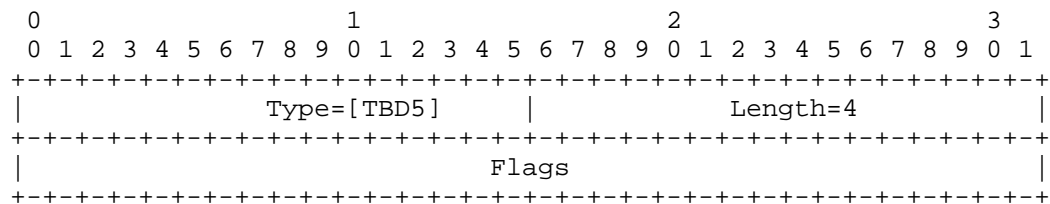
5.1. Capability Advertisement

During PCEP Initialization Phase, PCEP Speakers (PCE or PCC) advertise their support of Automatic Bandwidth Adjustment. A PCEP Speaker includes the "Auto-Bandwidth Capability" TLV, in the OPEN Object to advertise its support for PCEP Auto-Bandwidth extensions. The presence of the "Auto-Bandwidth Capability" TLV in the OPEN Object indicates that the Automatic Bandwidth Adjustment is supported as described in this document.

The PCEP protocol extensions for Auto-Bandwidth adjustments MUST NOT be used if one or both PCEP Speakers have not included the "Auto-Bandwidth Capability" TLV in their respective OPEN message. If the PCEP speaker that supports the extensions of this draft but did not advertise this capability, then upon receipt of AUTO-BANDWIDTH-ATTRIBUTE TLV in LSPA object, it SHOULD generate a PCErr with error-type 19 (Invalid Operation), error-value TBD4 (Auto-Bandwidth capability was not advertised) and it will terminate the PCEP session.

5.1.1 AUTO-BANDWIDTH-CAPABILITY TLV

The AUTO-BANDWIDTH-CAPABILITY TLV is an optional TLV for use in the OPEN Object for Automatic Bandwidth Adjustment via PCEP capability advertisement. Its format is shown in the following figure:



AUTO-BANDWIDTH-CAPABILITY TLV format

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

The value comprises a single field - Flags (32 bits). Currently no flags are defined for this TLV.

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

Advertisement of the Auto-Bandwidth capability TLV implies support of auto-bandwidth adjustment, as well as the objects, TLVs and procedures defined in this document.

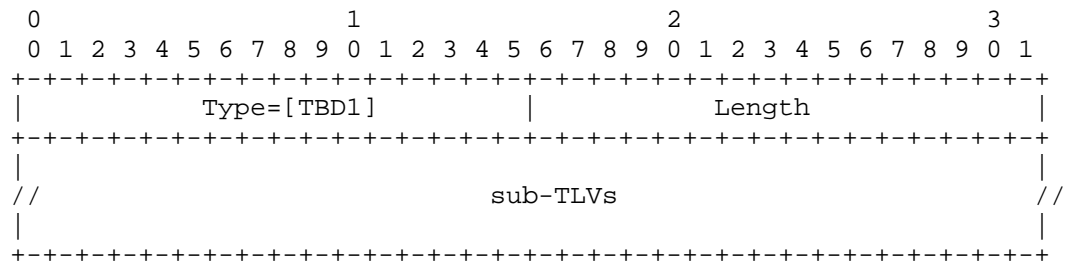
5.2. AUTO-BANDWIDTH-ATTRIBUTE TLV

The AUTO-BANDWIDTH-ATTRIBUTE TLV provides the 'configurable knobs' of the feature and it can be included as an optional TLV in the LSPA Object (as described in [RFC5440]).

For PCE-Initiated LSP ([I-D.ietf-pce-pce-initiated-lsp]), this TLV is included in the LSPA Object with PCInitiate message. For delegated LSPs, this TLV is carried in PCRpt message in LSPA Object.

The TLV is encoded in all PCEP messages for the LSP till the auto bandwidth adjustment feature is enabled, the absence of the TLV indicate the PCEP speaker wish to disable the feature.

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



AUTO-BANDWIDTH-ATTRIBUTE TLV format

Type: TBD1

Length: Variable

Value: This comprises one or more sub-TLVs.

Following sub-TLVs are defined in this document:

Type	Len	Name
1	4	Sample-Interval sub-TLV


```

2   4   Adjustment-Interval sub-TLV
3   4   Adjustment-Threshold sub-TLV
4   4   Adjustment-Threshold-Percentage sub-TLV
5   4   Minimum-Bandwidth sub-TLV
6   4   Maximum-Bandwidth sub-TLV
7   8   Overflow-Threshold sub-TLV
8   4   Overflow-Threshold-Percentage sub-TLV
9   8   Underflow-Threshold sub-TLV
10  4   Underflow-Threshold-Percentage sub-TLV

```

Future specification can define additional sub-TLVs.

The presence of AUTO-BANDWIDTH-ATTRIBUTE TLV in LSPA Object means that the automatic bandwidth adjustment feature is enabled. All sub-TLVs are optional and any unrecognized sub-TLV MUST be silently ignored. If a sub-TLV of same type appears more than once, only the first occurrence is processed and all others MUST be ignored.

The AUTO-BANDWIDTH-ATTRIBUTE TLV can also be carried in PCUpd message in LSPA Object in order to make updates to auto-bandwidth attributes such as Adjustment-Interval.

If sub-TLVs are not present, the default values based on the local policy are assumed.

The sub-TLVs are encoded to inform the PCEP peer the various sampling and adjustment parameters.

The following sub-sections describe the sub-TLVs which are currently defined to be carried within the AUTO-BANDWIDTH-ATTRIBUTE TLV.

5.2.1. Sample-Interval sub-TLV

The Sample-Interval sub-TLV specifies a time interval in seconds at which traffic samples are collected at the PCC.

The Type is 1, Length is 4, and the value comprises of 4-octet time interval, the valid range is from 1 to 604800, in seconds. The default value is 300 seconds.

```

      0               1               2               3
      0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
|                                     |                                     |
|      Type=1                         |      Length=4                     |
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
|                                     |                                     |
|      Sample-Interval                 |                                     |
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+

```

Sample-Interval sub-TLV format

5.2.2. Adjustment-Interval sub-TLV

The Adjustment-Interval sub-TLV specifies a time interval in seconds at which bandwidth adjustment should be made.

The Type is 2, Length is 4, and the value comprises of 4-octet time interval, the valid range is from 1 to 604800, in seconds. The default value is 300 seconds.

```

      0                   1                   2                   3
      0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-----+-----+-----+-----+-----+-----+-----+-----+
|                                     |                                     |
|               Type=2               |               Length=4           |
+-----+-----+-----+-----+-----+-----+-----+-----+
|                                     |                                     |
|               Adjustment-Interval   |                               |
+-----+-----+-----+-----+-----+-----+-----+-----+

```

Adjustment-Interval sub-TLV format

5.2.3. Adjustment Threshold

The sub-TLVs in this section are encoded to inform the PCEP peer the adjustment threshold parameters. An implementation MAY include both sub-TLVs for the absolute value and the percentage, in which case the bandwidth is adjusted when either of the adjustment threshold conditions are met.

5.2.3.1. Adjustment-Threshold sub-TLV

The Adjustment-Threshold sub-TLV is used to decide when the LSP bandwidth should be adjusted.

```

      0                   1                   2                   3
      0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-----+-----+-----+-----+-----+-----+-----+-----+
|                                     |                                     |
|               Type=3               |               Length=4           |
+-----+-----+-----+-----+-----+-----+-----+-----+
|                                     |                                     |
|               Adjustment Threshold   |                               |
+-----+-----+-----+-----+-----+-----+-----+-----+

```

Adjustment-Threshold sub-TLV format

The Type is 3, Length is 4, and the value comprises of -

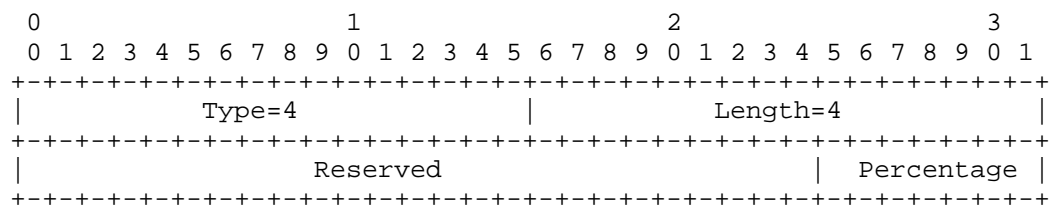
- o Adjustment Threshold: The absolute Adjustment-Threshold bandwidth value, encoded in IEEE floating point format (see

[IEEE.754.1985]), expressed in bytes per second. Refer to Section 3.1.2 of [RFC3471] for a table of commonly used values.

If the difference between the current MaxAvgBw and the current bandwidth reservation is greater than or equal to the threshold value, the LSP bandwidth is adjusted to the current bandwidth demand.

5.2.3.2. Adjustment-Threshold-Percentage sub-TLV

The Adjustment-Threshold-Percentage sub-TLV is used to decide when the LSP bandwidth should be adjusted.



Adjustment-Threshold-Percentage sub-TLV format

The Type is 4, Length is 4, and the value comprises of -

- o Reserved: SHOULD be set to zero on transmission and MUST be ignored on receipt.
- o Percentage: The Adjustment-Threshold value, encoded in percentage (an integer from 0 to 100). If the percentage difference between the current MaxAvgBw and the current bandwidth reservation is greater than or equal to the threshold percentage, the LSP bandwidth is adjusted to the current bandwidth demand.

5.2.4. Minimum and Maximum Bandwidth Values

5.2.4.1. Minimum-Bandwidth sub-TLV

The Minimum-Bandwidth sub-TLV specify the minimum bandwidth allowed for the LSP, and is expressed in bytes per second. The LSP bandwidth cannot be adjusted below the minimum bandwidth value.

The Type is 5, Length is 4, and the value comprises of 4-octet bandwidth value encoded in IEEE floating point format (see [IEEE.754.1985]), expressed in bytes per second. Refer to Section 3.1.2 of [RFC3471] for a table of commonly used values.

0	1	2	3
---	---	---	---

```

  0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-----+-----+-----+-----+-----+-----+-----+-----+
|               Type=5               |               Length=4       |
+-----+-----+-----+-----+-----+-----+-----+-----+
|               Minimum-Bandwidth     |
+-----+-----+-----+-----+-----+-----+-----+-----+

```

Minimum-Bandwidth sub-TLV format

5.2.4.2. Maximum-Bandwidth sub-TLV

The Maximum-Bandwidth sub-TLV specify the maximum bandwidth allowed for the LSP, and is expressed in bytes per second. The LSP bandwidth cannot be adjusted above the maximum bandwidth value.

The Type is 6, Length is 4, and the value comprises of 4-octet bandwidth value encoded in IEEE floating point format (see [IEEE.754.1985]), expressed in bytes per second. Refer to Section 3.1.2 of [RFC3471] for a table of commonly used values.

```

  0               1               2               3
  0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-----+-----+-----+-----+-----+-----+-----+-----+
|               Type=6               |               Length=4       |
+-----+-----+-----+-----+-----+-----+-----+-----+
|               Maximum-Bandwidth     |
+-----+-----+-----+-----+-----+-----+-----+-----+

```

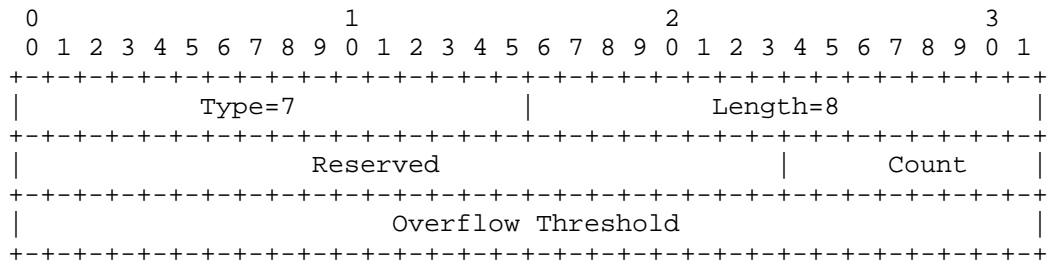
Maximum-Bandwidth sub-TLV format

5.2.5. Overflow and Underflow Conditions

The sub-TLVs in this section are encoded to inform the PCEP peer the overflow and underflow threshold parameters. An implementation MAY include sub-TLVs for the absolute value and the percentage for the threshold, in which case the bandwidth is immediately adjusted when either of the adjustment threshold conditions are met consecutively for the given count.

5.2.5.1. Overflow-Threshold sub-TLV

The Overflow-Threshold sub-TLV is used to decide if the bandwidth should be adjusted immediately.



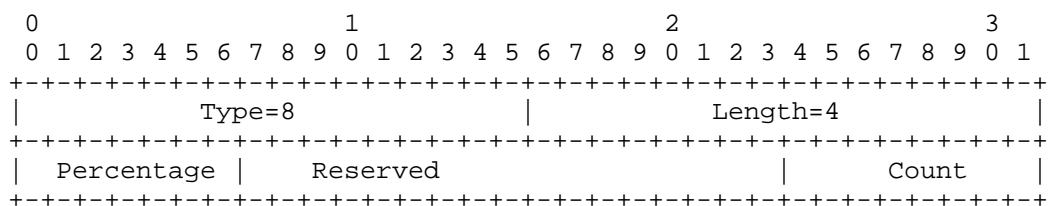
Overflow-Threshold sub-TLV format

The Type is 7, Length is 8, and the value comprises of -

- o Reserved: SHOULD be set to zero on transmission and MUST be ignored on receipt.
- o Count: The Overflow-Count value, encoded in integer. The value 0 is considered to be invalid. The number of consecutive samples for which the overflow condition MUST be met for the LSP bandwidth to be immediately adjusted to the current bandwidth demand, bypassing the adjustment-interval.
- o Overflow Threshold: The absolute Overflow-Threshold bandwidth value, encoded in IEEE floating point format (see [IEEE.754.1985]), expressed in bytes per second. Refer to Section 3.1.2 of [RFC3471] for a table of commonly used values. If the increase of the current MaxAvgBw from the current bandwidth reservation is greater than or equal to the threshold value, the overflow condition is met.

5.2.5.2. Overflow-Threshold-Percentage sub-TLV

The Overflow-Threshold-Percentage sub-TLV is used to decide if the bandwidth should be adjusted immediately.



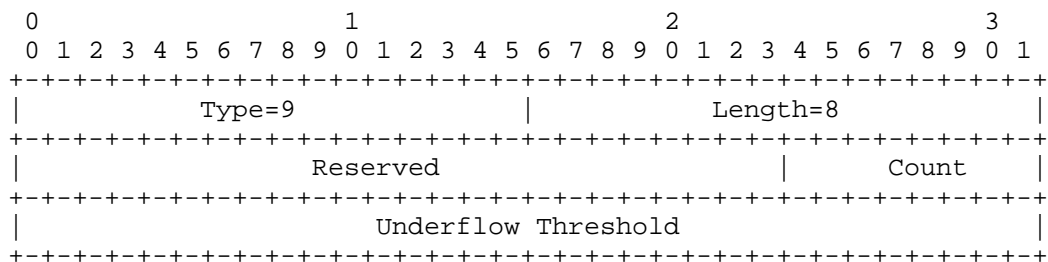
Overflow-Threshold-Percentage sub-TLV format

The Type is 8, Length is 4, and the value comprises of -

- o Percentage: The Overflow-Threshold value, encoded in percentage (an integer from 0 to 100). If the percentage increase of the current MaxAvgBw from the current bandwidth reservation is greater than or equal to the threshold percentage, the overflow condition is met.
- o Reserved: SHOULD be set to zero on transmission and MUST be ignored on receipt.
- o Count: The Overflow-Count value, encoded in integer. The value 0 is considered to be invalid. The number of consecutive samples for which the overflow condition MUST be met for the LSP bandwidth to be immediately adjusted to the current bandwidth demand, bypassing the adjustment-interval.

5.2.5.3. Underflow-Threshold sub-TLV

The Underflow-Threshold sub-TLV is used to decide if the bandwidth should be adjusted immediately.



Underflow-Threshold sub-TLV format

The Type is 9, Length is 8, and the value comprises of -

- o Reserved: SHOULD be set to zero on transmission and MUST be ignored on receipt.
- o Count: The Underflow-Count value, encoded in integer. The value 0 is considered to be invalid. The number of consecutive samples for which the underflow condition MUST be met for the LSP bandwidth to be immediately adjusted to the current bandwidth demand, bypassing the adjustment-interval.
- o Underflow Threshold: The absolute Underflow-Threshold bandwidth value, encoded in IEEE floating point format (see [IEEE.754.1985]), expressed in bytes per second. Refer to Section 3.1.2 of [RFC3471] for a table of commonly used values. If the decrease of the current MaxAvgBw from the current bandwidth

reservation is greater than or equal to the threshold value, the underflow condition is met.

5.2.5.4. Underflow-Threshold-Percentage sub-TLV

The Underflow-Threshold-Percentage sub-TLV is used to decide if the bandwidth should be adjusted immediately.

```

      0               1               2               3
      0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-----+-----+-----+-----+-----+-----+-----+-----+
|                                     |                                     |
|      Type=10                       |      Length=4                     |
+-----+-----+-----+-----+-----+-----+-----+-----+
|  Percentage  |  Reserved          |          Count          |
+-----+-----+-----+-----+-----+-----+-----+

```

Underflow-Threshold-Percentage sub-TLV format

The Type is 10, Length is 4, and the value comprises of -

- o Percentage: The Underflow-Threshold value, encoded in percentage (an integer from 0 to 100). If the percentage decrease of the current MaxAvgBw from the current bandwidth reservation is greater than or equal to the threshold percentage, the underflow condition is met.
- o Reserved: SHOULD be set to zero on transmission and MUST be ignored on receipt.
- o Count: The Underflow-Count value, encoded in integer. The value 0 is considered to be invalid. The number of consecutive samples for which the underflow condition MUST be met for the LSP bandwidth to be immediately adjusted to the current bandwidth demand, bypassing the adjustment-interval.

5.3. BANDWIDTH Object

As per [RFC5440], the BANDWIDTH object (Object-Class value 5) is defined with two Object-Type values as following:

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

PCC reports the calculated bandwidth to be adjusted (MaxAvgBw) to the

PCE using existing 'Requested Bandwidth with BANDWIDTH Object-Type as 1.

5.4. The PCInitiate Message

A PCInitiate message is a PCEP message sent by a PCE to a PCC to trigger LSP instantiation or deletion [I.D.ietf-pce-pce-initiated-lsp].

For the PCE-initiated LSP [I.D.ietf-pce-pce-initiated-lsp] with Auto-Bandwidth feature enabled, AUTO-BANDWIDTH-ATTRIBUTE TLV MUST be included in the LSPA object with the PCInitiate message. The rest of the processing remains unchanged.

5.5. The PCRpt Message

As specified in [I.D.ietf-pce-pce-initiated-lsp], the PCC creates the LSP using the attributes communicated by the PCE, and local values for the unspecified parameters. After the successful instantiation of the LSP, PCC automatically delegates the LSP to the PCE and generates an LSP State Report (PCRpt) for the LSP.

When LSP is delegated to a PCE for the very first time, BANDWIDTH object of type 1 is used to specify the requested bandwidth in the PCRpt message.

When the LSP is enabled with the Auto-Bandwidth feature, PCC SHOULD include the BANDWIDTH object of type 1 to specify the calculated bandwidth to be adjusted to the PCE in the PCRpt message.

The definition of the PCRpt message (see [I-D.ietf-pce-stateful-pce]) is unchanged by this document.

5.6. The PCNtf Message

As per [RFC5440], the PCEP Notification message (PCNtf) can be sent by a PCEP speaker to notify its peer of a specific event. As described in Section 4.3 of this document, a PCEP speaker SHOULD notify its PCEP peer that it is overwhelmed, and on receipt of such notification the peer SHOULD NOT send any PCEP messages related to auto-bandwidth adjustment. If a PCEP message related to auto-bandwidth adjustment is received, it MUST be silently ignored.

When a PCEP speaker is overwhelmed, it SHOULD notify its peer by sending a PCNtf message with Notification Type = TBD6 (Auto-bandwidth

Overwhelm State) and Notification Value = 1 (Entering auto-bandwidth overwhelm state). Optionally, OVERLOADED-DURATION TLV [RFC5440] MAY be included that specifies the time period during which no further PCEP messages related to auto-bandwidth adjustment should be sent. When the PCEP speaker is no longer in the overwhelm state and is available to process the auto-bandwidth adjustment, it SHOULD notify its peer by sending a PCNTf message with Notification Type = TBD6 (Auto-bandwidth Overwhelm State) and Notification Value = 2 (Clearing auto-bandwidth overwhelm state).

When Auto-Bandwidth feature is deployed, a PCE can send this notification to PCC when a PCC is reporting frequent auto-bandwidth adjustments. If a PCC is overwhelmed with re-signaling/re-routing, it can also notify the PCE to not adjust the LSP bandwidth while in overwhelm state.

6. Security Considerations

This document defines AUTO-BANDWIDTH-CAPABILITY TLV, AUTO-BANDWIDTH-ATTRIBUTE TLV which do not add any new security concerns beyond those discussed in [RFC5440] and [I-D.ietf-pce-stateful-pce].

Some deployments may find the reporting of the auto-bandwidth information as extra sensitive and thus SHOULD employ suitable PCEP security mechanisms like TCP-AO or [I-D.ietf-pce-pceps].

7. Manageability Considerations

7.1. Control of Function and Policy

The Auto-Bandwidth feature SHOULD be controlled per tunnel (at ingress (PCC) or PCE), the values for parameters like sample-interval, adjustment-interval, minimum-bandwidth, maximum-bandwidth, adjustment-threshold SHOULD be configurable by an operator.

7.2. Information and Data Models

[RFC7420] describes the PCEP MIB, there are no new MIB Objects defined in this document.

7.3. Liveness Detection and Monitoring

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

7.4. Verify Correct Operations

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

7.5. Requirements On Other Protocols

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

7.6. Impact On Network Operations

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

8. IANA Considerations

8.1. PCEP TLV Type Indicators

This document defines the following new PCEP TLVs; IANA is requested to make the following allocations from this registry.
<<http://www.iana.org/assignments/pcep/pcep.xhtml#pcep-tlv-type-indicators>>.

Value	Name	Reference
TBD5	AUTO-BANDWIDTH-CAPABILITY	[This I.D.]
TBD1	AUTO-BANDWIDTH-ATTRIBUTE	[This I.D.]

8.2. AUTO-BANDWIDTH-CAPABILITY TLV Flag Field

IANA is requested to create a registry to manage the Flag field of the AUTO-BANDWIDTH-CAPABILITY TLV.

New bit numbers may be allocated only by an IETF Consensus action. 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

No bit is defined for the AUTO-BANDWIDTH-CAPABILITY TLV Object flag field in this document.

8.3. AUTO-BANDWIDTH-ATTRIBUTE Sub-TLV

This document specifies the AUTO-BANDWIDTH-ATTRIBUTE Sub-TLVs. IANA is requested to create an "AUTO-BANDWIDTH-ATTRIBUTE Sub-TLV Types" sub-registry in the "PCEP TLV Type Indicators" for the sub-TLVs carried in the AUTO-BANDWIDTH-ATTRIBUTE TLV. This document defines the following types:

Type	Name	Reference
0	Reserved	[This I.D.]
1	Sample-Interval sub-TLV	[This I.D.]
2	Adjustment-Interval sub-TLV	[This I.D.]
3	Adjustment-Threshold sub-TLV	[This I.D.]
4	Adjustment-Threshold-Percentage sub-TLV	[This I.D.]

5	Minimum-Bandwidth sub-TLV	[This I.D.]
6	Maximum-Bandwidth sub-TLV	[This I.D.]
7	Overflow-Threshold sub-TLV	[This I.D.]
8	Overflow-Threshold-Percentage sub-TLV	[This I.D.]
9	Underflow-Threshold sub-TLV	[This I.D.]
10	Underflow-Threshold-Percentage sub-TLV	[This I.D.]
11-	Unassigned	[This I.D.]
65535		

8.4. Error Object

This document defines a new Error-Value for PCErr message of type 19 (Invalid Operation) [I-D.ietf-pce-stateful-pce]; IANA is requested to make the following allocation from this registry.
[<http://www.iana.org/assignments/pcep/pcep.xhtml#pcep-error-object>](http://www.iana.org/assignments/pcep/pcep.xhtml#pcep-error-object)

Error-Value	Meaning	Reference
TBD4	Auto-Bandwidth Capability was not Advertised	[This I.D.]

8.5. Notification Object

IANA is requested to allocate new Notification Types and Notification Values within the "Notification Object" sub-registry of the PCEP Numbers registry, as follows:

Type	Meaning	Reference
TBD6	Auto-Bandwidth Overwhelm State	[This I.D.]
	Notification-value=1: Entering Auto-Bandwidth overwhelm state	
	Notification-value=2: Clearing Auto-Bandwidth overwhelm state	

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.
- [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 (work in progress).

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

9.2. Informative References

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

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

[I-D.ietf-pce-stateful-pce-app] Zhang, X. and I. Minei, "Applicability of a Stateful Path Computation Element (PCE)", draft-ietf-pce-stateful-pce-app (work in progress).

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

[IEEE.754.1985] Institute of Electrical and Electronics Engineers, "Standard for Binary Floating-Point Arithmetic", IEEE Standard 754, August 1985.

Acknowledgments

Authors would like to thank Robert Varga, Venugopal Reddy, Reeja Paul, Sandeep Boina, Avantika, JP Vasseur and Himanshu Shah for their useful comments and suggestions.

Contributors' Addresses

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

EMail: kinghe@tencent.com

Xian Zhang
Huawei Technologies
Research Area F3-1B,
Huawei Industrial Base,
Shenzhen, 518129
China

Phone: +86-755-28972645
EMail: zhang.xian@huawei.com

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

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

Authors' Addresses

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

EMail: dhruv.ietf@gmail.com

Udayasree Palle
Huawei Technologies
Divyashree Techno Park, Whitefield
Bangalore, Karnataka 560037
India

EMail: udayasree.palle@huawei.com

Ravi Singh
Juniper Networks
1194 N. Mathilda Ave.
Sunnyvale, CA 94089
USA

EMail: ravis@juniper.net

Rakesh Gandhi
Individual Contributor

EMail: rgandhi.ietf@gmail.com

Luyuan Fang
eBay
USA

EMail: luyuanf@gmail.com

PCE Working Group
Internet-Draft
Intended status: Experimental
Expires: 6 September 2022

D. Dhody
S. Peng
Huawei Technologies
Y. Lee
Samsung Electronics
D. Ceccarelli
Ericsson
A. Wang
China Telecom
G. Mishra
Verizon Inc.
S. Sivabalan
Ciena Corporation
5 March 2022

PCEP extensions for Distribution of Link-State and TE Information
draft-dhodylee-pce-pcep-ls-23

Abstract

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

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

Requirements Language

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

Status of This Memo

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

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

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

This Internet-Draft will expire on 6 September 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 (<https://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 Revised BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Revised BSD License.

Table of Contents

1. Introduction	4
1.1. Scope	5
2. Terminology	6
3. Applicability	6
4. Requirements for PCEP extensions	7
5. New Functions to distribute link-state (and TE) via PCEP . .	8
6. Overview of Extensions to PCEP	9
6.1. New Messages	9
6.2. Capability Advertisement	9
6.3. Initial Link-State (and TE) Synchronization	10
6.3.1. Optimizations for LS Synchronization	12
6.4. LS Report	12
7. Transport	12
8. PCEP Messages	13
8.1. LS Report Message	13
8.2. The PCErr Message	13
9. Objects and TLV	14
9.1. TLV Format	14
9.2. Open Object	14
9.2.1. LS Capability TLV	14
9.3. LS Object	15
9.3.1. Routing Universe TLV	17
9.3.2. Route Distinguisher TLV	18
9.3.3. Virtual Network TLV	18
9.3.4. Local Node Descriptors TLV	18

9.3.5. Remote Node Descriptors TLV	19
9.3.6. Node Descriptors Sub-TLVs	20
9.3.7. Link Descriptors TLV	21
9.3.8. Prefix Descriptors TLV	21
9.3.9. PCEP-LS Attributes	22
9.3.9.1. Node Attributes TLV	22
9.3.9.2. Link Attributes TLV	22
9.3.9.3. Prefix Attributes TLV	23
9.3.10. Removal of an Attribute	23
10. Other Considerations	24
10.1. Inter-AS Links	24
11. Security Considerations	24
12. Manageability Considerations	24
12.1. Control of Function and Policy	24
12.2. Information and Data Models	25
12.3. Liveness Detection and Monitoring	25
12.4. Verify Correct Operations	25
12.5. Requirements On Other Protocols	26
12.6. Impact On Network Operations	26
13. IANA Considerations	26
13.1. PCEP Messages	26
13.2. PCEP Objects	26
13.3. LS Object	26
13.4. PCEP-Error Object	27
13.5. PCEP TLV Type Indicators	28
13.6. PCEP-LS Sub-TLV Type Indicators	28
14. TLV Code Points Summary	29
15. Implementation Status	30
15.1. Hierarchical Transport PCE controllers	30
15.2. ONOS-based Controller (MDSC and PNC)	31
16. Acknowledgments	31
17. References	31
17.1. Normative References	31
17.2. Informative References	32
Appendix A. Examples	35
A.1. All Nodes	35
A.2. Designated Node	37
A.3. Between PCEs	37
Appendix B. Contributor Addresses	38
Authors' Addresses	39

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 (Interior Gateway Protocol (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).

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 mechanism is offered by BGP-LS [I-D.ietf-idr-rfc7752bis] .

[RFC8231] describes a set of extensions to PCEP to provide stateful control. A stateful PCE has access to not only the information carried by the network's IGP, but also the set of active paths and their reserved resources for its computations. Path Computation Client (PCC) can delegate the rights to modify the LSP parameters to an Active Stateful PCE. This requires PCE to quickly be updated on any changes in the topology/TED, so that PCE can meet the need for updating LSPs effectively and in a timely manner. The fastest way for a PCE to be updated on TED changes is via a direct session with each network node and with an incremental update from each network node with only the attributes that gets modified.

[RFC8281] describes the setup, maintenance, and teardown of PCE-initiated LSPs under the stateful PCE model, without the need for local configuration on the PCC, thus allowing for a dynamic network that is centrally controlled and deployed. This model requires timely topology and TED update at the PCE.

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

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

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

Further [RFC6805] describes Hierarchical-PCE architecture, where a parent PCE maintains a domain topology map. To build this domain topology map, the child PCE can carry the border nodes and inter-domain link information to the parent PCE using the mechanism described in this document. Further as described in [RFC8637], the child PCE can also transport abstract Link-State and TE information from child PCE to a Parent PCE using the mechanism described in this document to build an abstract topology at the parent PCE.

[RFC8231] describe LSP state synchronization between PCCs and PCEs in case of stateful PCE. This document does not make any change to the LSP state synchronization process. The mechanism described in this document are on top of the existing LSP state synchronization.

1.1. Scope

The procedures described in this document are experimental. The experiment is intended to enable research for the usage of PCEP to populate the Link-State and TE Information from a PCC to the PCE. For this purpose, this document specifies new PCEP message and object/TLVs.

The new message introduced by this document will not be understood by legacy implementations. On receiving the message, a legacy implementation will behave according to the rules for a unknown message as per [RFC5440]. It is assumed that this experiment will be conducted only when both the PCE and PCC form part of the experiment.

It is possible that a PCC or PCE can operate with peers, some of which form part of the experiment and some that do not. In this case, the capability exchange required before using this extension would take care of the mismatch. A PCEP speaker that offers this feature to its peer that does not support or does not wish to support the feature will not receive indication of support in the Open message, and so is expected to not use the feature. Thus this experimentation would not clash with or cause harm to existing deployments. Further since a PCEP speaker would use the new message only after capability exchange, there is no danger of this experimentation "escaping" to the wider Internet. A PCEP speaker that receives the new message that is part of the feature when use of the feature has not been agreed, will send an error message as described in Section 6.9 of [RFC5440]. A PCEP speaker that receives the new object that is part of the feature when use of the feature has not been agreed, will send an error message as described in Section 7.2 of [RFC5440].

The experiment will end three years after the RFC is published. At that point, the RFC authors will attempt to determine how widely this has been implemented and deployed. When the results of implementation and deployment are available, this document (or part there of) will be updated and refined, and then it could be moved from Experimental to Standards Track.

2. Terminology

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

3. Applicability

The mechanism specified in this draft is applicable to deployments:

- * Where there is no IGP or BGP-LS running in the network.
- * Where there is no IGP or BGP-LS running at the PCE to learn link-state and TE information.
- * Where there is IGP or BGP-LS running but with a need for a faster and direct TE and link-state population and convergence at the PCE.
 - A PCE may receive partial information (say basic TE, link-state) from IGP and other information (optical and impairment) from PCEP.
 - A PCE may receive an incremental update (as opposed to the full (entire) information of the node/link).

- A PCE may receive full information from both existing mechanisms (IGP or BGP-LS) and PCEP.
- * Where there is a need for transporting (abstract) Link-State and TE information from child PCE to a Parent PCE in H-PCE [RFC6805]; as well as for Provisioning Network Controller (PNC) to Multi-Domain Service Coordinator (MDSC) in Abstraction and Control of TE Networks (ACTN) [RFC8453].
- * Where there is an existing PCEP session between all the nodes and the PCE-based central controller (PCECC) [RFC8283], and the operator would like to use PCEP as direct southbound interface to all the nodes in the network. This enables the operator to use PCEP as a single direct protocol between the controller and all the nodes in the network. In this mode, all nodes send only the local information.

Based on the local policy and deployment scenario, a PCC chooses to send only local information or both local and remote learned information. How a PCE manages the link-state (and TE) information is implementation specific and thus out of the scope of this document.

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

[RFC8453] defines three types of topology abstraction - (1) Native/White Topology; (2) Black Topology; and (3) Grey Topology. Based on the local policy, the PNC (or child PCE) would share the domain topology to the MDSC (or Parent PCE) based on the abstraction type. The protocol extensions defined in this document can carry any type of topology abstraction.

4. Requirements for PCEP extensions

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

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

2. PCC supporting this draft MUST have the capability to report the link-state (and TE) information to the PCE. This MUST include self originated (local) information and MAY also allow remote information learned via routing protocols. PCC MUST be capable to do the initial bulk sync at the time of session initialization as well as any changes there after.
 3. A PCE MAY learn link-state (and TE) from PCEP as well as from existing mechanisms like IGP/BGP-LS. PCEP extensions MUST have a mechanism to correlate the information learned via other means. There MUST NOT be any changes to the existing link-state (and TE) population mechanism via IGP/BGP-LS. PCEP extension SHOULD keep the properties in a protocol (IGP or BGP-LS) neutral way, such that an implementation need not know about any OSPF or IS-IS or BGP-LS protocol specifics.
 4. It SHOULD be possible to encode only the changes in link-state (and TE) properties (after the initial sync) in PCEP messages. This leads to faster convergence.
 5. The same mechanism SHOULD be used for both MPLS TE as well as GMPLS, optical, and impairment aware properties.
 6. The same mechanism SHOULD be used for PCE to PCE Link-state (and TE) synchronization.
5. New Functions to distribute link-state (and TE) via PCEP

Several new functions are required in PCEP to support distribution of link-state (and TE) information. A function can be initiated either from a PCC towards a PCE (C-E) or from a PCE towards a PCC (E-C). The new functions are:

- * Capability advertisement (E-C,C-E): both the PCC and the PCE MUST announce during PCEP session establishment that they support PCEP extensions for distribution of link-state (and TE) information defined in this document.
- * Link-State (and TE) synchronization (C-E): after the session between the PCC and a PCE is initialized, the PCE must learn Link-State (and TE) information before it can perform path computations. In the case of stateful PCE it is RECOMMENDED that this operation be done before LSP state synchronization.
- * Link-State (and TE) Report (C-E): a PCC sends an LS (and TE) report to a PCE whenever the Link-State and TE information changes.

6. Overview of Extensions to PCEP

6.1. New Messages

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

6.2. Capability Advertisement

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

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

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

6.3. Initial Link-State (and TE) Synchronization

The purpose of LS Synchronization is to provide a checkpoint-in-time state replica of a PCC's link-state (and TE) database in a PCE. State Synchronization is performed immediately after the Initialization phase (see [RFC5440]). In case of stateful PCE ([RFC8231]) it is RECOMMENDED that the LS synchronization should be done before LSP state synchronization.

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

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

If the PCC encounters a problem which prevents it from completing the LS synchronization, it MUST send a PCErr message with error-type TBD2 (LS Synchronization Error) and error-value 2 (indicating an internal PCC error) to the PCE and terminate the session.

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

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

The successful LS Synchronization sequence is shown in Figure 1.

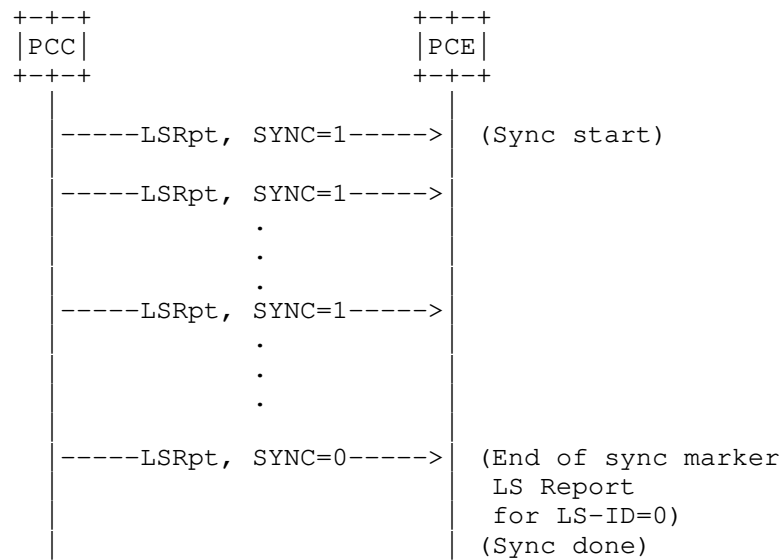


Figure 1: Successful LS synchronization

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

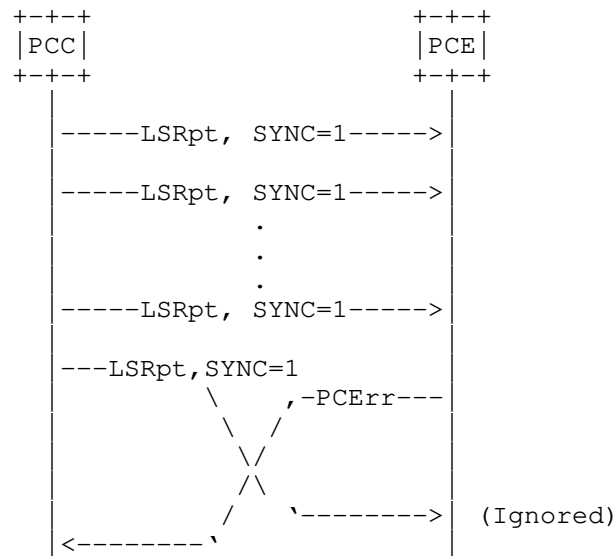


Figure 2: Failed LS synchronization (PCE failure)

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

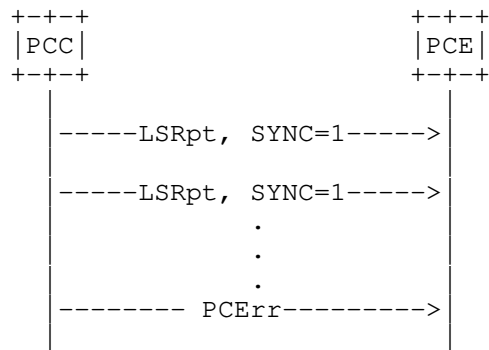


Figure 3: Failed LS synchronization (PCC failure)

6.3.1. Optimizations for LS Synchronization

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

6.4. LS Report

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

More details about the LSRpt message are in Section 8.1.

7. Transport

A permanent PCEP session (section 4.2.8 of [RFC5440]) MUST be established between a PCE and PCC supporting link-state (and TE) distribution via PCEP. In the case of session failure, session re-establishment is re-attempted as per the procedures defined in [RFC5440].

8. PCEP Messages

As defined in [RFC5440], a PCEP message consists of a common header followed by a variable-length body made of a set of objects that can be either mandatory or optional. An object is said to be mandatory in a PCEP message when the object must be included for the message to be considered valid. For each PCEP message type, a set of rules is defined that specify the set of objects that the message can carry. An implementation MUST form the PCEP messages using the object ordering specified in this document.

8.1. LS Report Message

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

The format of the LSRpt message is as follows:

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

Where:

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

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

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

8.2. The PCErr Message

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

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

```

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

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

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

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

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

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

```

9. Objects and TLV

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

9.1. TLV Format

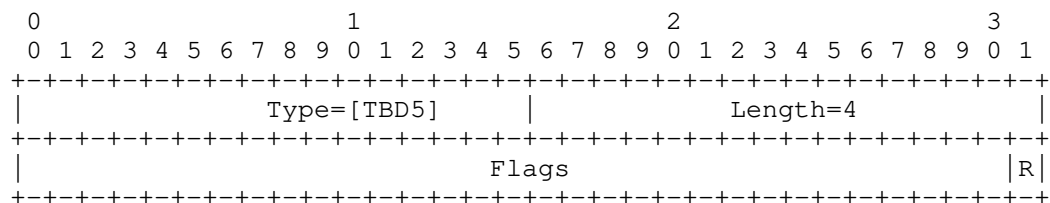
The TLV and the sub-TLV format (and padding) in this document, is as per section 7.1 of [RFC5440].

9.2. Open Object

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

9.2.1. LS Capability TLV

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



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

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

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

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

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

9.3. LS Object

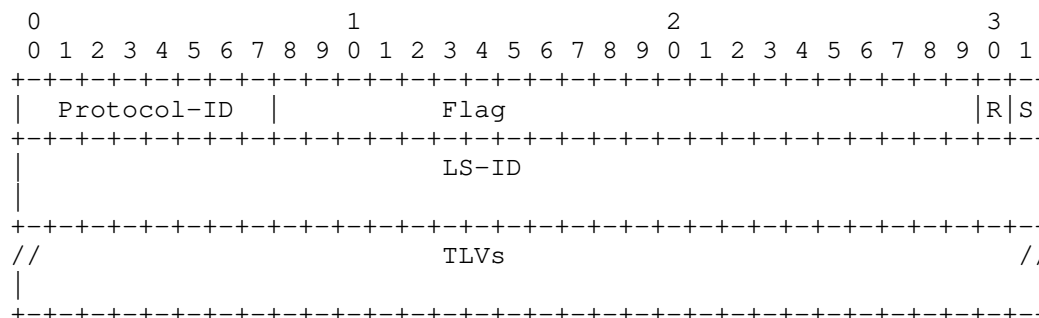
The LS (link-state) object MUST be carried within LSRpt messages and MAY be carried within PCErr messages. The LS object contains a set of fields used to specify the target node or link. It also contains a flag indicating to a PCE that the LS synchronization is in progress. The TLVs used with the LS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

LS Object-Class is TBD6.

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

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

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



Protocol-ID (8-bit): The field provides the source information. The protocol could be an IGP, BGP-LS, or an abstraction algorithm. In case PCC only provides local information of the PCC, it MUST use Protocol-ID as Direct. The following values are defined (some of the initial values are the same as [I-D.ietf-idr-rfc7752bis]):

Protocol-ID	Source protocol
1	IS-IS Level 1
2	IS-IS Level 2
3	OSPFv2
4	Direct
5	Static configuration
6	OSPFv3
7	BGP
8	RSVP-TE
9	Segment Routing
10	PCEP
11	Abstraction

Flags (24-bit):

- * S (SYNC - 1 bit): the S Flag MUST be set to 1 on each LSRpt sent from a PCC during LS Synchronization. The S Flag MUST be set to 0 in other LSRpt messages sent from the PCC.
- * R (Remove - 1 bit): On LSRpt messages, the R Flag indicates that the node/link/prefix has been removed from the PCC and the PCE SHOULD remove from its database. Upon receiving an LS Report with the R Flag set to 1, the PCE SHOULD remove all state for the node/link/prefix identified by the LS Identifiers from its database.

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

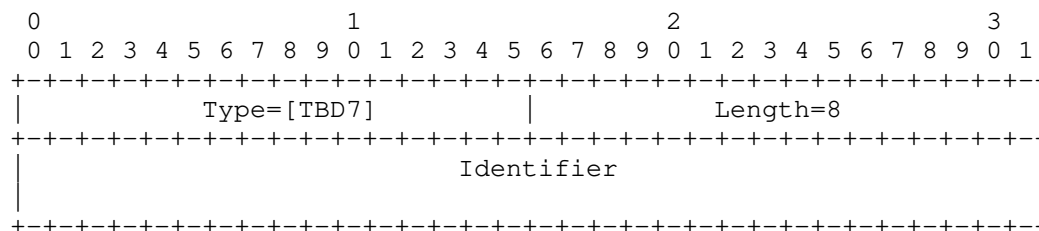
Unassigned bits are considered reserved. They MUST be set to 0 on transmission and MUST be ignored on receipt.

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

9.3.1. Routing Universe TLV

In the case of remote link-state (and TE) population when existing IGP/BGP-LS are also used, OSPF and IS-IS may run multiple routing protocol instances over the same link as described in [I-D.ietf-idr-rfc7752bis]. See [RFC8202] and [RFC6549] for more information. These instances define an independent "routing universe". The 64-bit 'Identifier' field is used to identify the "routing universe" where the LS object belongs. The LS objects representing IGP objects (nodes or links or prefix) from the same routing universe MUST have the same 'Identifier' value; LS objects with different 'Identifier' values MUST be considered to be from different routing universes.

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



The below table lists the 'Identifier' values that are defined as well-known in this draft (same as [I-D.ietf-idr-rfc7752bis]).

Identifier	Routing Universe
0	Default Layer 3 Routing topology

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

9.3.2. Route Distinguisher TLV

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

The ROUTE-DISTINGUISHER TLV is defined in [RFC9168] as a Flow Specification TLVs with a separate registry. This document also adds the ROUTE-DISTINGUISHER TLV with TBD15 in the PCEP TLV registry to be used inside the LS object.

9.3.3. Virtual Network TLV

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

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

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

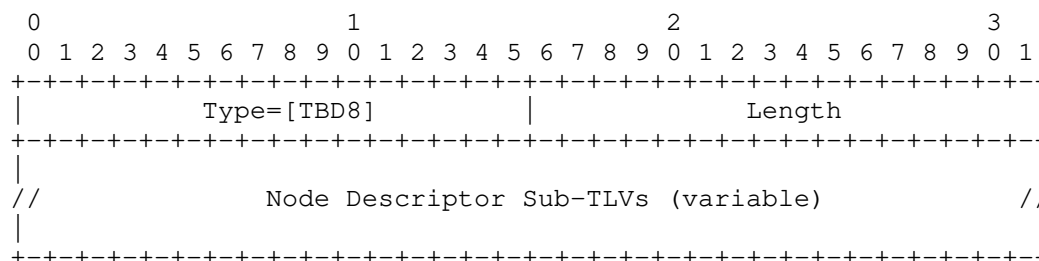
9.3.4. Local Node Descriptors TLV

As described in [I-D.ietf-idr-rfc7752bis], each link is anchored by a pair of Router-IDs that are used by the underlying IGP, namely, 48-bit ISO System-ID for IS-IS and 32-bit Router-ID for OSPFv2 and OSPFv3. In case of additional auxiliary Router-IDs used for TE, these MUST also be included in the link attribute TLV (see Section 9.3.9.2).

It is desirable that the Router-ID assignments inside the Node Descriptors TLV are globally unique. Some considerations for globally unique Node/Link/Prefix identifiers are described in [I-D.ietf-idr-rfc7752bis].

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

This TLV is applicable for all LS Object-Type.

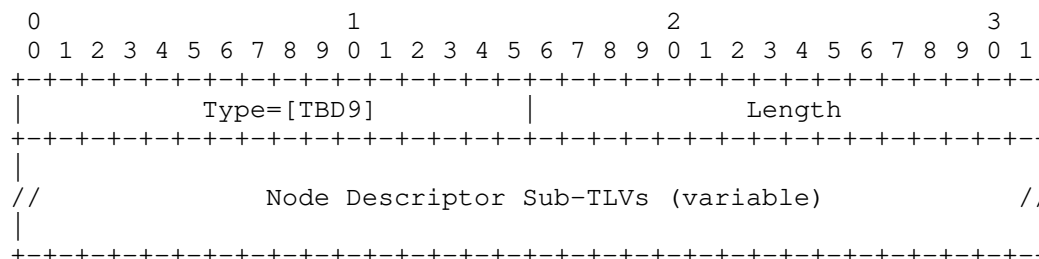


The value contains one or more Node Descriptor Sub-TLVs defined in Section 9.3.6.

9.3.5. Remote Node Descriptors TLV

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

This TLV is applicable for LS Link Object-Type.



9.3.6. Node Descriptors Sub-TLVs

The Node Descriptors TLV (Local and Remote) carries one or more Node Descriptor Sub-TLV follows the format of all PCEP TLVs as defined in [RFC5440], however, the Type values are selected from a new PCEP-LS sub-TLV IANA registry (see Section 13.6).

Type values are chosen so that there can be commonality with BGP-LS [I-D.ietf-idr-rfc7752bis]. This is possible because the "BGP-LS Node Descriptor, Link Descriptor, Prefix Descriptor, and Attribute TLVs" registry marks 0-255 as reserved. Thus the space of the sub-TLV values for the Type field can be partitioned as shown below -

Range	
0	Reserved - must not be allocated.
1 .. 255	New PCEP sub-TLV allocated according to the registry defined in this document.
256 .. 65535	Per BGP registry defined by [I-D.ietf-idr-rfc7752bis]. Not to be allocated in this registry.

All Node Descriptors TLVs defined for BGP-LS can then be used with PCEP-LS as well. One new PCEP sub-TLVs for Node Descriptor are defined in this document.

Sub-TLV	Description	Length	Value defined in
1	SPEAKER-ENTITY-ID	Variable	[RFC8232]

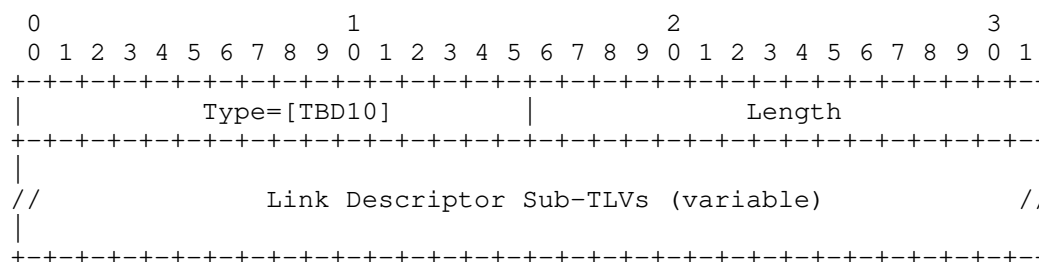
A new sub-TLV type (1) is allocated for SPEAKER-ENTITY-ID sub-TLV. The length and value fields are as per [RFC8232].

9.3.7. Link Descriptors TLV

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

The 'Link descriptor' TLVs uniquely identify a link among multiple parallel links between a pair of anchor routers similar to [I-D.ietf-idr-rfc7752bis].

This TLV is applicable for LS Link Object-Type.



All Link Descriptors TLVs defined for BGP-LS can then be used with PCEP-LS as well. No new PCEP sub-TLVs for Link Descriptor are defined in this document.

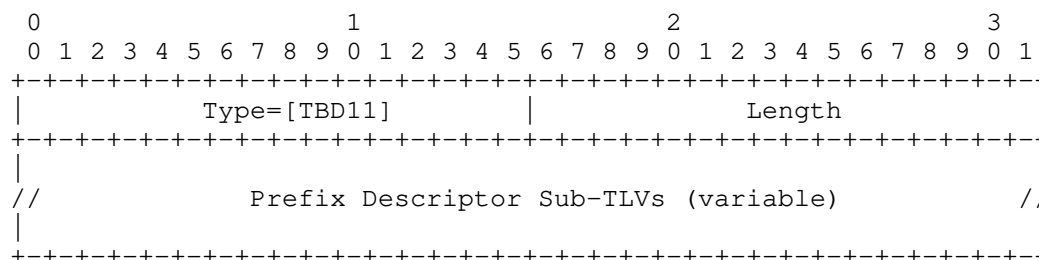
The format and semantics of the 'value' fields in most 'Link Descriptor' sub-TLVs correspond to the format and semantics of value fields in IS-IS Extended IS Reachability sub-TLVs, defined in [RFC5305], [RFC5307] and [RFC6119]. Although the encodings for 'Link Descriptor' TLVs were originally defined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

The information about a link present in the LSA/LSP originated by the local node of the link determines the set of sub-TLVs in the Link Descriptor of the link as described in [I-D.ietf-idr-rfc7752bis].

9.3.8. Prefix Descriptors TLV

The Prefix Descriptors TLV contains Prefix Descriptors that uniquely identify an IPv4 or IPv6 Prefix originated by a Node. This TLV MUST be included in the LS Report when during a given PCEP session a prefix is first reported to a PCE. A PCC sends to a PCE the first LS Report either during State Synchronization, or when a new prefix is learned at the PCC. The length of this TLV is variable.

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

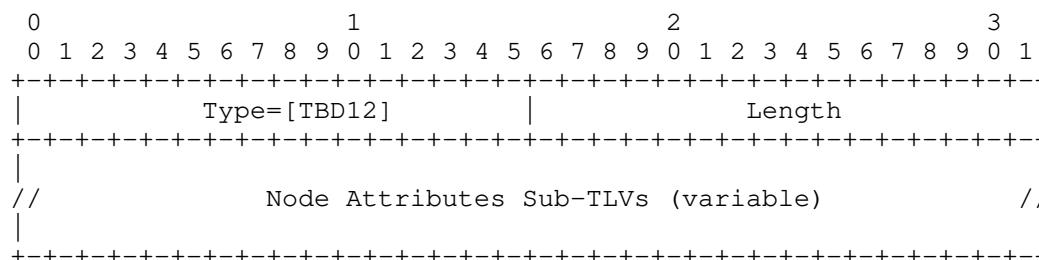


All Prefix Descriptors TLVs defined for BGP-LS can then be used with PCEP-LS as well. No new PCEP sub-TLVs for Prefix Descriptor are defined in this document.

9.3.9. PCEP-LS Attributes

9.3.9.1. Node Attributes TLV

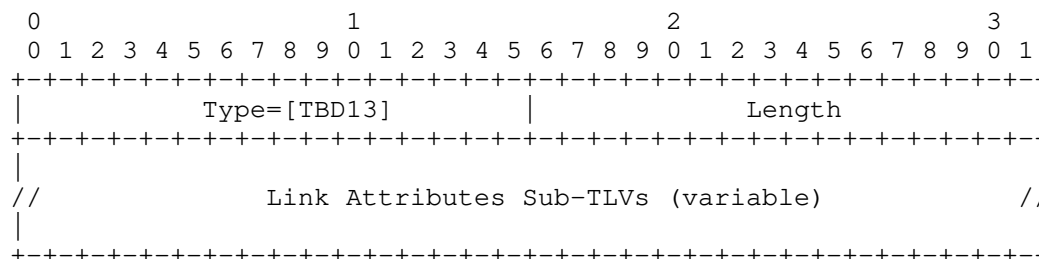
This is an optional attribute that is used to carry node attributes. This TLV is applicable for LS Node Object-Type.



All Node Attributes TLVs defined for BGP-LS can then be used with PCEP-LS as well. No new PCEP sub-TLVs for Node Attributes are defined in this document.

9.3.9.2. Link Attributes TLV

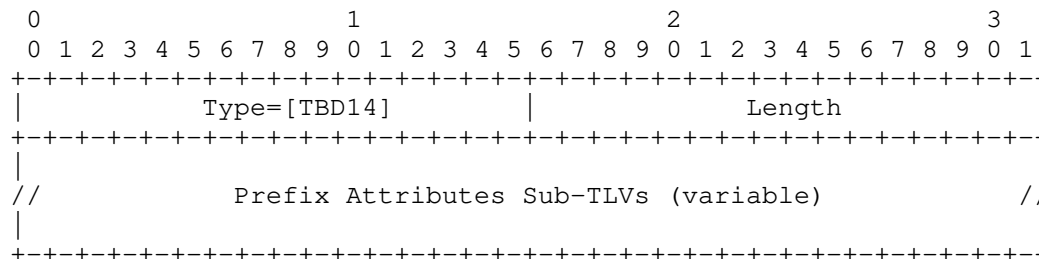
This TLV is applicable for LS Link Object-Type. The format and semantics of the 'value' fields in some 'Link Attribute' sub-TLVs correspond to the format and semantics of the 'value' fields in IS-IS Extended IS Reachability sub-TLVs, defined in [RFC5305], [RFC5307] and [I-D.ietf-idr-rfc7752bis]. Although the encodings for 'Link Attribute' TLVs were originally defined for IS-IS, the TLVs can carry data sourced either by IS-IS or OSPF or direct.



All Link Attributes TLVs defined for BGP-LS can then be used with PCEP-LS as well. No new PCEP sub-TLVs for Link Attributes are defined in this document.

9.3.9.3. Prefix Attributes TLV

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



All Prefix Attributes TLVs defined for BGP-LS can then be used with PCEP-LS as well. No new PCEP sub-TLVs for Prefix Attributes are defined in this document.

9.3.10. Removal of an Attribute

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

10. Other Considerations

10.1. Inter-AS Links

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

11. Security Considerations

This document extends PCEP for LS (and TE) distribution including a new LSRpt message with a new object and TLVs. Procedures and protocol extensions defined in this document do not effect the overall PCEP security model. See [RFC5440], [RFC8253]. Tampering with the LSRpt message may have an effect on path computations at PCE. It also provides adversaries an opportunity to eavesdrop and learn sensitive information and plan sophisticated attacks on the network infrastructure. The PCE implementation SHOULD provide mechanisms to prevent strains created by network flaps and amount of LS (and TE) information. Thus it is suggested that any mechanism used for securing the transmission of other PCEP message be applied here as well. As a general precaution, it is RECOMMENDED that these PCEP extensions only are activated on authenticated and encrypted sessions belonging to the same administrative authority.

Further, as stated in [RFC6952], PCEP implementations SHOULD support the TCP-AO [RFC5925] and not use TCP MD5 because of TCP MD5's known vulnerabilities and weaknesses. PCEP also support Transport Layer Security (TLS) [RFC8253] as per the recommendations and best current practices in [RFC7525].

12. Manageability Considerations

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

12.1. Control of Function and Policy

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

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

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

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

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

12.2. Information and Data Models

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

An implementation SHOULD also provide the statistics:

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

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

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

12.3. Liveness Detection and Monitoring

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

12.4. Verify Correct Operations

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

12.5. Requirements On Other Protocols

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

12.6. Impact On Network Operations

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

13. IANA Considerations

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

13.1. PCEP Messages

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

Value	Meaning	Reference
TBD3	LSRpt	[This I-D]

13.2. PCEP Objects

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

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

13.3. LS Object

This document requests that a new sub-registry, named "LS Object Protocol-ID Field", is created within the "Path Computation Element Protocol (PCEP) Numbers" registry to manage the Flag field of the LSP object. New values are to be assigned by Standards Action [RFC8126].

Value	Meaning	Reference
0	Reserved	[This I-D]
1	IS-IS Level 1	[This I-D]
2	IS-IS Level 2	[This I-D]
3	OSPFv2	[This I-D]
4	Direct	[This I-D]
5	Static configuration	[This I-D]
6	OSPFv3	[This I-D]
7	BGP	[This I-D]
8	RSVP-TE	[This I-D]
9	Segment Routing	[This I-D]
10	PCEP	[This I-D]
11	Abstraction	[This I-D]
12-255	Unassigned	

Further, this document also requests that a new sub-registry, named "LS Object Flag Field", is created within the "Path Computation Element Protocol (PCEP) Numbers" registry to manage the Flag field of the LSP object. New values are to be assigned by Standards Action [RFC8126]. Each bit should be tracked with the following qualities:

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

The following values are defined in this document:

Bit	Description	Reference
0-21	Unassigned	
22	R (Remove bit)	[This I-D]
23	S (Sync bit)	[This I-D]

13.4. PCEP-Error Object

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

Error-Type	Meaning	Reference
6	Mandatory Object missing Error-Value=TBD4 (LS object missing)	[RFC5440] [This I-D]
19	Invalid Operation Error-Value=TBD1 (Attempted LS Report if LS remote capability was not advertised)	[RFC8231] [This I-D]
TBD2	LS Synchronization Error Error-Value=1 (An error in processing the LSRpt) Error-Value=2 (An internal PCC error)	[This I-D]

13.5. PCEP TLV Type Indicators

This document defines the following new PCEP TLVs.

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

13.6. PCEP-LS Sub-TLV Type Indicators

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

Allocations from this registry are to be made according to the following assignment policies [RFC8126]:

Range	Assignment policy
0	Reserved - must not be allocated.
1 .. 251	Specification Required
252 .. 255	Experimental Use
256 .. 65535	Reserved - must not be allocated. Usage mirrors the BGP-LS TLV registry [I-D.ietf-idr-rfc7752bis]

IANA is requested to pre-populate this registry with values defined in this document as follows, taking the new values from the range 1 to 251:

Value	Meaning
1	SPEAKER-ENTITY-ID

14. TLV Code Points Summary

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

TLV	Description	Ref TLV	Value defined in:
TBD7	Routing Universe	--	Sec 9.2.1
TBD15	Route Distinguisher	--	Sec 9.2.2
*	Virtual Network	--	[ietf-pce-vn-association]
TBD8	Local Node Descriptors	256	[I-D.ietf-idr-rfc7752bis] /3.2.1.2
TBD9	Remote Node Descriptors	257	[I-D.ietf-idr-rfc7752bis] /3.2.1.3
TBD10	Link Descriptors	--	Sec 9.2.8
TBD11	Prefix Descriptors	--	Sec 9.2.9
TBD12	Node Attributes	--	Sec 9.2.10.1
TBD13	Link Attributes	--	Sec 9.2.10.2
TBD14	Prefix Attributes	--	Sec 9.2.10.3

* this TLV is defined in a different PCEP document

Figure 4: TLV Table

15. Implementation Status

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

15.1. Hierarchical Transport PCE controllers

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

- connectivity services on the ACTN based recursive hierarchical SDN/PCE platform that has the three-tier level SDN controllers (two-tier level MDSC and PNC) on the top of the PTN systems managed by EMS.
- Integration test of two tier-level MDSC: The SBI of the low level MDSC is the YANG based Korean national standards and the one of the high-level MDSC the PCEP-LS based ACTN protocols.

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

15.2. ONOS-based Controller (MDSC and PNC)

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

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

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

Further description can be found at ONOS-PCEP (<https://wiki.onosproject.org/display/ONOS/PCEP+Protocol>) and the code at ONOS-PCEP-GITHUB (<https://github.com/opennetworkinglab/onos/tree/master/protocols/pcep>).

16. Acknowledgments

This document borrows some of the structure and text from the [I-D.ietf-idr-rfc7752bis].

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

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

17. References

17.1. Normative References

[I-D.ietf-idr-rfc7752bis]
Talaulikar, K., "Distribution of Link-State and Traffic Engineering Information Using BGP", Work in Progress, Internet-Draft, draft-ietf-idr-rfc7752bis-10, 10 November 2021, <<https://datatracker.ietf.org/doc/html/draft-ietf-idr-rfc7752bis-10>>.

- [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>>.
- [RFC5305] Li, T. and H. Smit, "IS-IS Extensions for Traffic Engineering", RFC 5305, DOI 10.17487/RFC5305, October 2008, <<https://www.rfc-editor.org/info/rfc5305>>.
- [RFC5307] Kompella, K., Ed. and Y. Rekhter, Ed., "IS-IS Extensions in Support of Generalized Multi-Protocol Label Switching (GMPLS)", RFC 5307, DOI 10.17487/RFC5307, October 2008, <<https://www.rfc-editor.org/info/rfc5307>>.
- [RFC5440] Vasseur, JP., Ed. and JL. Le Roux, Ed., "Path Computation Element (PCE) Communication Protocol (PCEP)", RFC 5440, DOI 10.17487/RFC5440, March 2009, <<https://www.rfc-editor.org/info/rfc5440>>.
- [RFC6119] Harrison, J., Berger, J., and M. Bartlett, "IPv6 Traffic Engineering in IS-IS", RFC 6119, DOI 10.17487/RFC6119, February 2011, <<https://www.rfc-editor.org/info/rfc6119>>.
- [RFC8174] Leiba, B., "Ambiguity of Uppercase vs Lowercase in RFC 2119 Key Words", BCP 14, RFC 8174, DOI 10.17487/RFC8174, May 2017, <<https://www.rfc-editor.org/info/rfc8174>>.
- [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>>.

17.2. Informative References

- [I-D.ietf-pce-pcep-yang]
Dhody, D., Hardwick, J., Beeram, V. P., and J. Tantsura, "A YANG Data Model for Path Computation Element Communications Protocol (PCEP)", Work in Progress, Internet-Draft, draft-ietf-pce-pcep-yang-18, 25 January 2022, <<https://datatracker.ietf.org/doc/html/draft-ietf-pce-pcep-yang-18>>.
- [I-D.ietf-pce-vn-association]
Lee, Y., Zheng, H., and D. Ceccarelli, "Path Computation Element communication Protocol (PCEP) extensions for Establishing Relationships between sets of LSPs and Virtual Networks", Work in Progress, Internet-Draft,

draft-ietf-pce-vn-association-05, 15 October 2021,
<<https://datatracker.ietf.org/doc/html/draft-ietf-pce-vn-association-05>>.

[I-D.ietf-teas-actn-requirements]

Lee, Y., Ceccarelli, D., Miyasaka, T., Shin, J. Y., and K. Lee, "Requirements for Abstraction and Control of TE Networks", Work in Progress, Internet-Draft, draft-ietf-teas-actn-requirements-09, 2 March 2018,
<<https://datatracker.ietf.org/doc/html/draft-ietf-teas-actn-requirements-09>>.

[I-D.kondreddy-pce-pcep-ls-sync-optimizations]

Kondreddy, V. R. and M. S. Negi, "Optimizations of PCEP Link-State (LS) Synchronization Procedures", Work in Progress, Internet-Draft, draft-kondreddy-pce-pcep-ls-sync-optimizations-00, 9 October 2015,
<<https://datatracker.ietf.org/doc/html/draft-kondreddy-pce-pcep-ls-sync-optimizations-00>>.

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

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

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

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

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

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

- [RFC5925] Touch, J., Mankin, A., and R. Bonica, "The TCP Authentication Option", RFC 5925, DOI 10.17487/RFC5925, June 2010, <<https://www.rfc-editor.org/info/rfc5925>>.
- [RFC6549] Lindem, A., Roy, A., and S. Mirtorabi, "OSPFv2 Multi-Instance Extensions", RFC 6549, DOI 10.17487/RFC6549, March 2012, <<https://www.rfc-editor.org/info/rfc6549>>.
- [RFC6805] King, D., Ed. and A. Farrel, Ed., "The Application of the Path Computation Element Architecture to the Determination of a Sequence of Domains in MPLS and GMPLS", RFC 6805, DOI 10.17487/RFC6805, November 2012, <<https://www.rfc-editor.org/info/rfc6805>>.
- [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, DOI 10.17487/RFC6952, May 2013, <<https://www.rfc-editor.org/info/rfc6952>>.
- [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, <<https://www.rfc-editor.org/info/rfc7525>>.
- [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>>.
- [RFC8202] Ginsberg, L., Previdi, S., and W. Henderickx, "IS-IS Multi-Instance", RFC 8202, DOI 10.17487/RFC8202, June 2017, <<https://www.rfc-editor.org/info/rfc8202>>.
- [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>>.
- [RFC8253] Lopez, D., Gonzalez de Dios, O., Wu, Q., and D. Dhody, "PCEPS: Usage of TLS to Provide a Secure Transport for the Path Computation Element Communication Protocol (PCEP)", RFC 8253, DOI 10.17487/RFC8253, October 2017, <<https://www.rfc-editor.org/info/rfc8253>>.

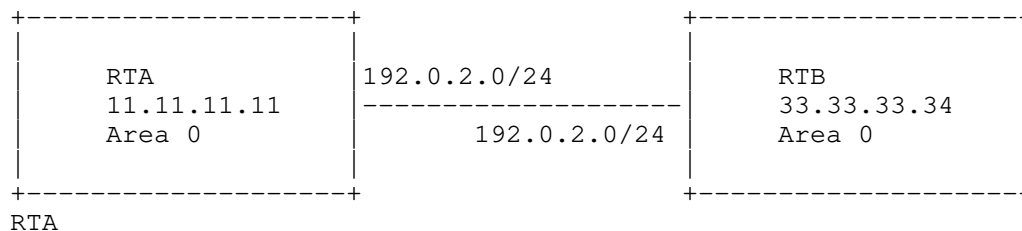
- [RFC8281] Crabbe, E., Minei, I., Sivabalan, S., and R. Varga, "Path Computation Element Communication Protocol (PCEP) Extensions for PCE-Initiated LSP Setup in a Stateful PCE Model", RFC 8281, DOI 10.17487/RFC8281, December 2017, <<https://www.rfc-editor.org/info/rfc8281>>.
- [RFC8283] Farrel, A., Ed., Zhao, Q., Ed., Li, Z., and C. Zhou, "An Architecture for Use of PCE and the PCE Communication Protocol (PCEP) in a Network with Central Control", RFC 8283, DOI 10.17487/RFC8283, December 2017, <<https://www.rfc-editor.org/info/rfc8283>>.
- [RFC8453] Ceccarelli, D., Ed. and Y. Lee, Ed., "Framework for Abstraction and Control of TE Networks (ACTN)", RFC 8453, DOI 10.17487/RFC8453, August 2018, <<https://www.rfc-editor.org/info/rfc8453>>.
- [RFC8637] Dhody, D., Lee, Y., and D. Ceccarelli, "Applicability of the Path Computation Element (PCE) to the Abstraction and Control of TE Networks (ACTN)", RFC 8637, DOI 10.17487/RFC8637, July 2019, <<https://www.rfc-editor.org/info/rfc8637>>.
- [RFC9168] Dhody, D., Farrel, A., and Z. Li, "Path Computation Element Communication Protocol (PCEP) Extension for Flow Specification", RFC 9168, DOI 10.17487/RFC9168, January 2022, <<https://www.rfc-editor.org/info/rfc9168>>.

Appendix A. Examples

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

A.1. All Nodes

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



LS Node

TLV - Local Node Descriptors
 Sub-TLV - 514: OSPF Area-ID: 0.0.0.0
 Sub-TLV - 515: IGP Router-ID: 11.11.11.11
TLV - Node Attributes TLV
 Sub-TLV(s)

LS Link

TLV - Local Node Descriptors
 Sub-TLV - 514: OSPF Area-ID: 0.0.0.0
 Sub-TLV - 515: IGP Router-ID: 11.11.11.11
TLV - Remote Node Descriptors
 Sub-TLV - 514: OSPF Area-ID: 0.0.0.0
 Sub-TLV - 515: IGP Router-ID: 22.22.22.22
TLV - Link Descriptors
 Sub-TLV - 259: IPv4 interface: 192.0.2.1
 Sub-TLV - 260: IPv4 neighbor: 192.0.2.2
TLV - Link Attributes TLV
 Sub-TLV(s)

RTB

LS Node

TLV - Local Node Descriptors
 Sub-TLV - 514: OSPF Area-ID: 0.0.0.0
 Sub-TLV - 515: IGP Router-ID: 22.22.22.22
TLV - Node Attributes TLV
 Sub-TLV(s)

LS Link

TLV - Local Node Descriptors
 Sub-TLV - 514: OSPF Area-ID: 0.0.0.0
 Sub-TLV - 515: IGP Router-ID: 22.22.22.22
TLV - Remote Node Descriptors
 Sub-TLV - 514: OSPF Area-ID: 0.0.0.0
 Sub-TLV - 515: IGP Router-ID: 11.11.11.11
TLV - Link Descriptors
 Sub-TLV - 259: IPv4 interface: 192.0.2.2
 Sub-TLV - 260: IPv4 neighbor: 192.0.2.1
TLV - Link Attributes TLV
 Sub-TLV(s)

A similar example with IPv6 address (say 2001:db8::1 and 2001:db8::2) for the links could be imagined with all other information as same and just IPv6 interface and neighbor TLVs.

A.2. Designated Node

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

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

A.3. Between PCEs

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

PCE1

LS Node

TLV - Local Node Descriptors

Sub-TLV - 512: Autonomous System: 100 (Domain 1)

Sub-TLV - 515: IGP Router-ID: 11.11.11.11 (abstract)

LS Link

TLV - Local Node Descriptors

Sub-TLV - 512: Autonomous System: 100

Sub-TLV - 515: IGP Router-ID: 11.11.11.11 (abstract)

TLV - Remote Node Descriptors

Sub-TLV - 512: Autonomous System: 200 (Domain 2)

Sub-TLV - 515: IGP Router-ID: 22.22.22.22 (abstract)

TLV - Link Descriptors

Sub-TLV - 259: IPv4 interface: 192.0.2.1

Sub-TLV - 260: IPv4 neighbor: 192.0.2.2

TLV - Link Attributes TLV

Sub-TLV(s)

LS Link

TLV - Local Node Descriptors

Sub-TLV - 512: Autonomous System: 100

Sub-TLV - 515: IGP Router-ID: 11.11.11.11 (abstract)

TLV - Remote Node Descriptors

Sub-TLV - 512: Autonomous System: 200

Sub-TLV - 515: IGP Router-ID: 22.22.22.22 (abstract)

TLV - Link Descriptors

Sub-TLV - 259: IPv4 interface: 198.51.100.1

Sub-TLV - 260: IPv4 neighbor: 198.51.100.2

TLV - Link Attributes TLV
Sub-TLV(s)

LS Link

TLV - Local Node Descriptors
Sub-TLV - 512: Autonomous System: 100
Sub-TLV - 515: IGP Router-ID: 11.11.11.11 (abstract)
TLV - Remote Node Descriptors
Sub-TLV - 512: Autonomous System: 400 (Domain 4)
Sub-TLV - 515: IGP Router-ID: 44.44.44.44 (abstract)
TLV - Link Descriptors
Sub-TLV - 259: IPv4 interface: 203.0.113.1
Sub-TLV - 260: IPv4 neighbor: 203.0.113.2
TLV - Link Attributes TLV
Sub-TLV(s)

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

Further the exact border nodes and abstract internal path between the border nodes may also be transported to the Parent PCE to enable ACTN as described in [RFC8637] using the similar LS node and link objects encodings.

Appendix B. Contributor Addresses

Udayasree Palle

Email: udayasreereddy@gmail.com

Sergio Belotti
Nokia

Email: sergio.belotti@nokia.com

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

Email: satishk@huawei.com

Cheng Li
Huawei Technologies
Huawei Campus, No. 156 Beiqing Rd.
Beijing 100095
China

Email: c.l@huawei.com

Authors' Addresses

Dhruv Dhody
Huawei Technologies
Divyashree Techno Park, Whitefield
Bangalore 560066
Karnataka
India
Email: dhruv.ietf@gmail.com

Shuping Peng
Huawei Technologies
Huawei Bld., No.156 Beiqing Rd.
Beijing
100095
China
Email: pengshuping@huawei.com

Young Lee
Samsung Electronics
Seoul
South Korea
Email: younglee.tx@gmail.com

Daniele Ceccarelli
Ericsson
Torshamnsgatan, 48
Stockholm
Sweden
Email: daniele.ceccarelli@ericsson.com

Aijun Wang
China Telecom
Beiqijia Town, Changping District
Beijing
102209
China
Email: wangaijun@tsinghua.org.cn

Gyan Mishra
Verizon Inc.
Email: gyan.s.mishra@verizon.com

Siva Sivabalan
Ciena Corporation
Email: ssivabal@ciena.com

Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: March 4, 2016

J. Dong
M. Chen
D. Dhody
Huawei Technologies
J. Tantsura
Ericsson
September 1, 2015

BGP Extensions for Path Computation Element (PCE) Discovery
draft-dong-pce-discovery-proto-bgp-03

Abstract

In networks where Path Computation Element (PCE) is used for centralized path computation, it is desirable for Path Computation Clients (PCCs) to automatically discover a set of PCEs and select the suitable ones to establish the PCEP session. RFC 5088 and RFC 5089 define the PCE discovery mechanisms based on Interior Gateway Protocols (IGP). This document describes several scenarios in which the IGP based PCE discovery mechanisms cannot be used directly. This document specifies the BGP extensions for PCE discovery in these scenarios. The BGP based PCE discovery mechanism is complementary to the existing IGP based mechanisms.

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 March 4, 2016.

Copyright Notice

Copyright (c) 2015 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	4
2.1. PCE Address Information	4
2.2. PCE Discovery TLVs	5
3. Operational Considerations	6
4. IANA Considerations	7
5. Security Considerations	7
6. Acknowledgements	7
7. References	7
7.1. Normative References	7
7.2. Informative References	8
Authors' Addresses	9

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 a set of PCEs and select the suitable ones to establish the PCEP session. [RFC5088] and [RFC5089] define the PCE discovery mechanisms based on Interior Gateway Protocols (IGP). Those IGP based mechanisms may not work in several scenarios where the PCEs do not participate in the IGP, and it is difficult for PCEs to participate in multiple IGP domains where PCE discovery is needed.

In some scenarios, Backward Recursive Path Computation (BRPC) [RFC5441] can be used by cooperating PCEs to compute inter-domain path, in which case these cooperating PCEs should be known to each other in advance. In case of inter-AS networks where the PCEs do not participate in a common IGP, the existing IGP discovery mechanism cannot be used to discover the PCEs in other domains.

In the Hierarchical PCE scenario [RFC6805], the child PCEs need to know the address of the parent PCEs. This cannot be achieved through IGP based discovery, as normally the child PCEs and the parent PCE are under different administration and reside in different domains.

Besides, as BGP could be used for north-bound distribution of routing and Label Switched Path (LSP) information to PCE as described in [I-D.ietf-idr-ls-distribution] [I-D.ietf-idr-te-lsp-distribution] and [I-D.ietf-idr-te-pm-bgp], PCEs can obtain the routing information without participating in IGP. In this scenario, some other PCE discovery mechanism is needed.

A detailed set of requirements for a PCE discovery mechanism are provided in [RFC4674].

This document proposes to extend BGP for PCE discovery in the above scenarios. In networks where BGP-LS is used for the north-bound routing information distribution to PCE, the BGP based PCE discovery can reuse the existing BGP sessions and mechanisms to achieve PCE discovery. It should be noted that 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.

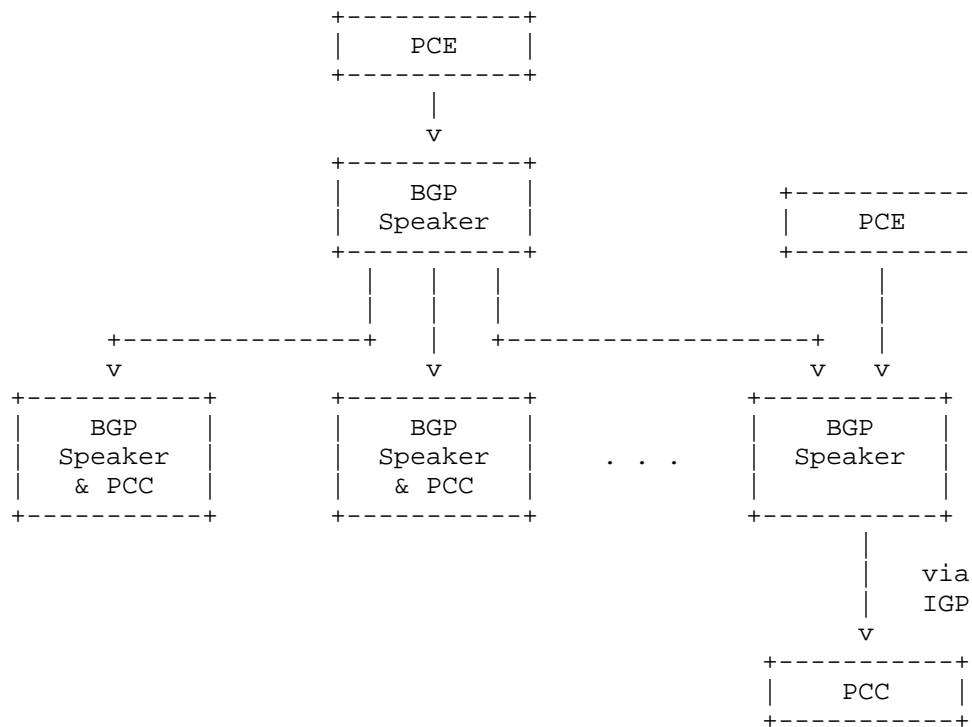


Figure 1: BGP for PCE discovery

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 collected from the network elements and distributed to PCE, while the PCE discovery information is advertised from PCE to PCCs, or between different PCEs. The PCCs maybe co-located with the BGP speakers as shown in Figure 1. The IGP based PCE discovery mechanism may be used for the distribution of PCE discovery information in 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]. The AFI and SAFI defined in [I-D.ietf-idr-ls-distribution] are re-used, and a new NLRI Type is defined for PCE discovery information as below:

- o Type = TBD: PCE Discovery NLRI

The format of PCE Discovery NLRI is shown in the following figure:

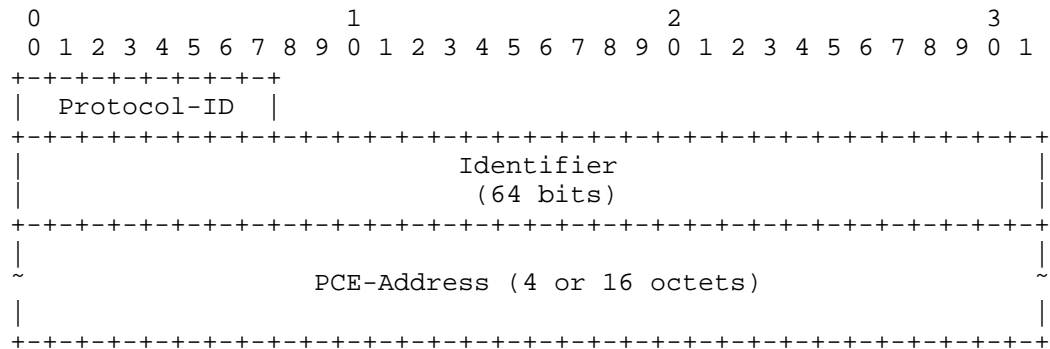
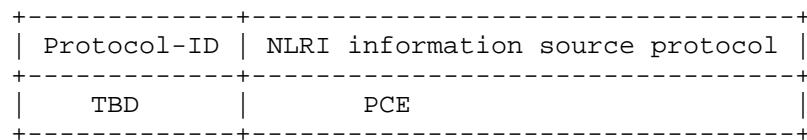


Figure 2. PCE Discovery NLRI

The 'Protocol-ID' field SHOULD be set to a new value which indicates the information source protocol is PCE.



As defined in [I-D.ietf-idr-ls-distribution], the 64-Bit 'Identifier' field is used to identify the "routing universe" where the PCE belongs.

2.2. PCE Discovery TLVs

The detailed PCE discovery information is carried in the BGP-LS attribute [I-D.ietf-idr-ls-distribution] with a new "PCE Discovery TLV", which contains a set of sub-TLVs for specific PCE discovery information. The PCE Discovery TLV and sub-TLVs SHOULD only be used with the PCE Discovery NLRI.

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

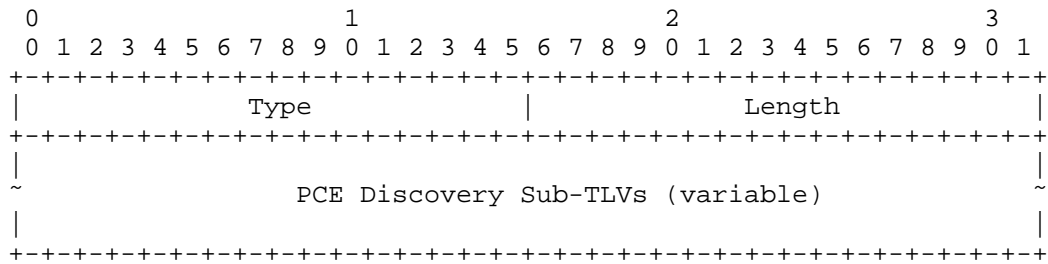


Figure 3. PCE Discovery TLV

The PCE Discovery sub-TLVs are listed as below. The format of the PCE Discovery sub-TLVs are consistent with the IGP PCED sub-TLVs as defined in [RFC5088] and [RFC5089]. The PATH-SCOPE sub-TLV MUST always be carried in the PCE Discovery TLV. Other PCE Discovery sub-TLVs are optional and may facilitate the PCE selection process on the PCCs.

Type	Length	Name
TBD	3	PATH-SCOPE sub-TLV
TBD	variable	PCE-CAP-FLAGS sub-TLV
TBD	variable	OSPF-PCE-DOMAIN sub-TLV
TBD	variable	IS-IS-PCE-DOMAIN sub-TLV
TBD	variable	OSPF-NEIG-PCE-DOMAIN sub-TLV
TBD	variable	IS-IS-NEIG-PCE-DOMAIN sub-TLV

More PCE Discovery sub-TLVs may be defined in future and the format SHOULD be in line with the new sub-TLVs defined for IGP based PCE discovery.

3. Operational Considerations

Existing BGP operational procedures apply to the advertisement of PCE discovery information. This information is treated as pure application level data which has no immediate impact on forwarding states. Normal BGP path selection can be applied to PCE Discovery NLRI only for the information propagation in the network, while the PCE selection on the PCCs would be based on the information carried in the PCE Discovery TLV.

The 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 a new NLRI Type for 'PCE Discovery NLRI' from the "BGP-LS NLRI-Types" registry.

IANA needs to assign a new Protocol-ID for "PCE" from the "BGP-LS Protocol-IDs" registry.

IANA needs to assign a new TLV code point for 'PCE Discovery TLV' from the "node anchor, link descriptor and link attribute TLVs" registry.

IANA needs to create a new registry for "PCE Discovery Sub-TLVs". The registry will be initialized as shown in section 2.2 of this document.

5. Security Considerations

Procedures and protocol extensions defined in this document do not affect the BGP security model. See the 'Security Considerations' section of [RFC4271] for a discussion of BGP security. Also refer to [RFC4272] and [RFC6952] for analysis of security issues for BGP.

6. Acknowledgements

The authors would like to thank Zhenbin Li and Hannes Gredler for their discussion and comments.

7. References

7.1. Normative References

- [I-D.ietf-idr-ls-distribution]
Gredler, H., Medved, J., Previdi, S., Farrel, A., and S. Ray, "North-Bound Distribution of Link-State and TE Information using BGP", draft-ietf-idr-ls-distribution-11 (work in progress), June 2015.
- [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>>.
- [RFC4271] Rekhter, Y., Ed., Li, T., Ed., and S. Hares, Ed., "A Border Gateway Protocol 4 (BGP-4)", RFC 4271, DOI 10.17487/RFC4271, January 2006, <<http://www.rfc-editor.org/info/rfc4271>>.

- [RFC4760] Bates, T., Chandra, R., Katz, D., and Y. Rekhter, "Multiprotocol Extensions for BGP-4", RFC 4760, DOI 10.17487/RFC4760, January 2007, <<http://www.rfc-editor.org/info/rfc4760>>.
- [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>>.

7.2. Informative References

- [I-D.ietf-idr-te-lsp-distribution]
Dong, J., Chen, M., Gredler, H., Previdi, S., and J. Tantsura, "Distribution of MPLS Traffic Engineering (TE) LSP State using BGP", draft-ietf-idr-te-lsp-distribution-03 (work in progress), May 2015.
- [I-D.ietf-idr-te-pm-bgp]
Wu, Q., Previdi, S., Gredler, H., Ray, S., and J. Tantsura, "BGP attribute for North-Bound Distribution of Traffic Engineering (TE) performance Metrics", draft-ietf-idr-te-pm-bgp-02 (work in progress), January 2015.
- [RFC4272] Murphy, S., "BGP Security Vulnerabilities Analysis", RFC 4272, DOI 10.17487/RFC4272, January 2006, <<http://www.rfc-editor.org/info/rfc4272>>.
- [RFC4674] Le Roux, J., Ed., "Requirements for Path Computation Element (PCE) Discovery", RFC 4674, DOI 10.17487/RFC4674, October 2006, <<http://www.rfc-editor.org/info/rfc4674>>.
- [RFC5441] Vasseur, JP., Ed., Zhang, R., Bitar, N., and JL. Le Roux, "A Backward-Recursive PCE-Based Computation (BRPC) Procedure to Compute Shortest Constrained Inter-Domain Traffic Engineering Label Switched Paths", RFC 5441, DOI 10.17487/RFC5441, April 2009, <<http://www.rfc-editor.org/info/rfc5441>>.

- [RFC6805] King, D., Ed. and A. Farrel, Ed., "The Application of the Path Computation Element Architecture to the Determination of a Sequence of Domains in MPLS and GMPLS", RFC 6805, DOI 10.17487/RFC6805, November 2012, <<http://www.rfc-editor.org/info/rfc6805>>.
- [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, DOI 10.17487/RFC6952, May 2013, <<http://www.rfc-editor.org/info/rfc6952>>.

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

Dhruv Dhody
Huawei Technologies
Leela Palace
Bangalore, Karnataka 560008
India

Email: dhruv.ietf@gmail.com

Jeff Tantsura
Ericsson
300 Holger Way
San Jose, CA 95134
US

Email: jeff.tantsura@ericsson.com

Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: January 4, 2018

J. Dong
M. Chen
D. Dhody
Huawei Technologies
J. Tantsura
Individual
K. Kumaki
KDDI Corporation
T. Murai
Furukawa Network Solution Corp.
July 3, 2017

BGP Extensions for Path Computation Element (PCE) Discovery
draft-dong-pce-discovery-proto-bgp-07

Abstract

In networks where a Path Computation Element (PCE) is used for path computation, it is desirable for the Path Computation Clients (PCCs) to discover dynamically and automatically a set of PCEs along with certain information relevant for PCE selection. RFC 5088 and RFC 5089 define the PCE discovery mechanisms based on Interior Gateway Protocols (IGP). This document defines extensions to BGP for the advertisement of PCE Discovery information. The BGP based PCE discovery mechanism is complementary to the existing IGP based mechanisms.

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 4, 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	2
2. Carrying PCE Discovery Information in BGP	3
2.1. PCE NLRI	3
2.1.1. PCE Descriptors	4
2.2. PCE Attribute TLVs	5
2.2.1. PCE Domain TLV	6
2.2.2. Neighbor PCE Domain TLV	6
3. Operational Considerations	7
4. IANA Considerations	7
5. Security Considerations	7
6. Contributors	7
7. Acknowledgements	8
8. References	8
8.1. Normative References	8
8.2. Informative References	9
Authors' Addresses	9

1. Introduction

In networks where a Path Computation Element (PCE) is used for path computation, it is desirable for the Path Computation Clients (PCCs) to discover dynamically and automatically a set of PCEs along with certain information relevant for PCE selection. [RFC5088] and [RFC5089] define the PCE discovery mechanisms based on Interior Gateway Protocols (IGP). 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.

[RFC4674] presents a set of requirements for a PCE discovery mechanism. This includes the discovery by a PCC of a set of one or more PCEs which may potentially be in some other domains. This is a desirable function in the case of inter-domain path computation. For example, Backward Recursive Path Computation (BRPC) [RFC5441] can be used by cooperating PCEs to compute an inter-AS path, in which case the discovery of PCE as well as the domain information is useful.

BGP has been extended for north-bound distribution of routing and TE information to PCE [RFC7752] and [I-D.ietf-idr-te-pm-bgp]. Similarly this document extends BGP to also carry the PCE discovery information.

This document defines extensions to BGP to allow a PCE to advertise its location, along with some information useful to a PCC for the PCE selection, so as to satisfy dynamic PCE discovery requirements set forth in [RFC4674].

This specification contains two parts: definition of a new BGP-LS NLRI [RFC7752] that describes PCE information and definition of PCE Attribute TLVs as part of BGP-LS attributes.

2. Carrying PCE Discovery Information in BGP

2.1. PCE NLRI

The PCE discovery information is advertised in BGP UPDATE messages using the MP_REACH_NLRI and MP_UNREACH_NLRI attributes [RFC4760]. The "Link- State NLRI" defined in [RFC7752] is extended to carry the PCE information. BGP speakers that wish to exchange PCE discovery information MUST use the BGP Multiprotocol Extensions Capability Code (1) to advertise the corresponding (AFI, SAFI) pair, as specified in [RFC4760].

The format of "Link-State NLRI" is defined in [RFC7752]. A new "NLRI Type" is defined for PCE Information as following:

- o Type = TBD1: PCE NLRI

The format of PCE NLRI is shown in the following figure:

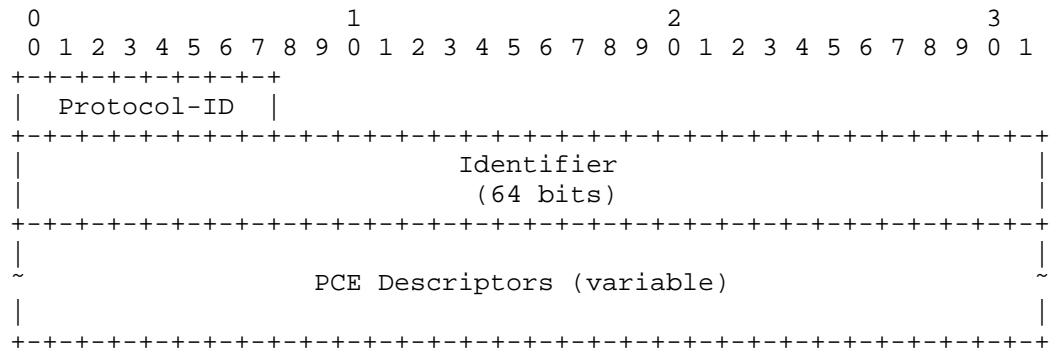


Figure 1. PCE NLRI

The 'Protocol-ID' field is defined in [RFC7752], to be set to the appropriate value that indicates the source of the PCE information. If BGP speaker and PCE are co-located, the Protocol-ID SHOULD be set to "Direct". If PCE information to advertise is configured at the BGP speaker, the Protocol-ID SHOULD be set to "Static configuration".

As defined in [RFC7752], the 64-Bit 'Identifier' field is used to identify the "routing universe" where the PCE belongs.

2.1.1.1. PCE Descriptors

The PCE Descriptor field is a set of Type/Length/Value (TLV) triplets. The format of each TLV is as per Section 3.1 of [RFC7752]. The PCE Descriptor TLVs uniquely identify a PCE. The following PCE descriptor are defined -

Codepoint	Descriptor TLV	Length
TBD2	IPv4 PCE Address	4
TBD3	IPv6 PCE Address	16

Table 1: PCE Descriptors

The PCE address TLVs specifies an IP address that can be used to reach the PCE. The PCE-ADDRESS Sub-TLV defined in [RFC5088] and [RFC5089] is used in the OSPF and IS-IS respectively. The format of the PCE address TLV are -

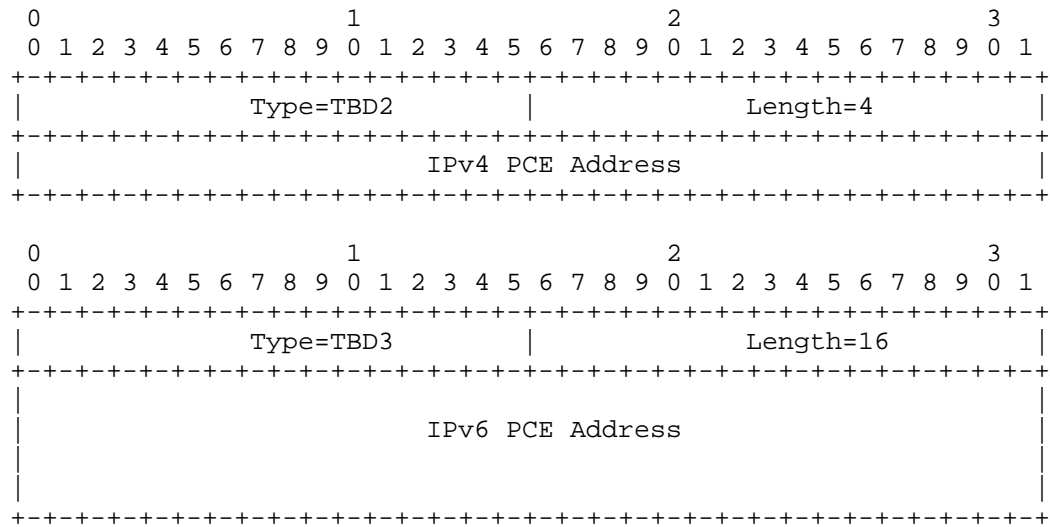


Figure 2. PCE Address TLVs

When the PCE has both an IPv4 and IPv6 address, both the TLVs MAY be included.

2.2. PCE Attribute TLVs

PCE Attribute TLVs are TLVs that may be encoded in the BGP-LS attribute [RFC7752] with a PCE NLRI. The format of each TLV is as per Section 3.1 of [RFC7752]. The format and semantics of the Value fields in some PCE Attribute TLVs correspond to the format and semantics of the Value fields in IS-IS PCED Sub-TLV, defined in [RFC5089]. Other PCE Attribute TLVs are defined in this document.

The following PCE Attribute TLVs are valid in the BGP-LS attribute with a PCE NLRI:

TLV Code Point	Description	IS-IS TLV /Sub-TLV	Reference (RFC/Section)
TBD4	Path Scope	5/2	[RFC5089]/4.2
TBD5	PCE Domain	-	-
TBD6	Neighbor PCE Domain	-	-
TBD7	PCE Capability	5/5	[RFC5089]/4.5

Table 2: PCE Attribute TLVs

The format and semantics of Path Scope and PCE capability is as per [RFC5089]. The Path Scope TLV is mandatory.

2.2.1. PCE Domain TLV

The PCE Domain TLV specifies a PCE-Domain (IGP area and/or AS) where the PCE has topology visibility and through which the PCE can compute paths.

```

      0               1               2               3
    0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-----+-----+-----+-----+-----+-----+-----+-----+
|                                     |                                     |
|                               Type=TBD5                               |
|                                     |                                     |
+-----+-----+-----+-----+-----+-----+-----+-----+
|                                     |                                     |
|                               Domain Sub-TLVs (variable)             |
|                                     |                                     |
+-----+-----+-----+-----+-----+-----+-----+-----+

```

The length of this TLV is variable. The value contains one or more domain sub-TLVs as listed below -

Sub-TLV Code Point	Description	Length
512	Autonomous System	4
514	OSPF Area-ID	4
1027	IS-IS Area Identifier	Variable

Multiple sub-TLVs MAY be included, when the PCE has visibility into multiple PCE-Domains.

2.2.2. Neighbor PCE Domain TLV

The Neighbor PCE Domain TLV specifies a neighbor PCE-Domain (IGP area and/or AS) toward which a PCE can compute paths.

```

      0               1               2               3
    0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-----+-----+-----+-----+-----+-----+-----+-----+
|                                     |                                     |
|                               Type=TBD6                               |
|                                     |                                     |
+-----+-----+-----+-----+-----+-----+-----+-----+
|                                     |                                     |
|                               Domain Sub-TLVs (variable)             |
|                                     |                                     |
+-----+-----+-----+-----+-----+-----+-----+-----+

```

The length of this TLV is variable. The value contains one or more domain sub-TLVs as listed above. Multiple sub-TLVs MAY be included, when the PCE can compute paths towards several neighbor PCE-Domains.

3. Operational Considerations

Existing BGP-LS operational procedures apply to the advertisement of PCE information as per [RFC7752]. This information is treated as pure application level data which has no immediate impact on forwarding states. The PCE information SHOULD be advertised only to the domains where such information is allowed to be used. This can be achieved by policy control on the ASBRs.

The PCE 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 a new NLRI Type for 'PCE NLRI' from the "BGP-LS NLRI-Types" registry.

IANA needs to assign new TLV code point as per Table 1 and 2 from the "BGP-LS Node Descriptor, Link Descriptor, Prefix Descriptor, and Attribute TLVs" registry.

[Editor's Note - Check if name of the registry should be changes with following instructions - Further IANA is requested to rename the registry as "BGP-LS Node Descriptor, Link Descriptor, Prefix Descriptor, PCE Descriptor, and Attribute TLVs".]

5. Security Considerations

Procedures and protocol extensions defined in this document do not affect the BGP security model. See the 'Security Considerations' section of [RFC4271] for a discussion of BGP security. Also refer to [RFC4272] and [RFC6952] for analysis of security issues for BGP.

Existing BGP-LS security considerations as per [RFC7752] continue to apply.

6. Contributors

The following individuals gave significant contributions to this document:

Takuya Miyasaka
KDDI Corporation
ta-miyasaka@kddi.com

7. Acknowledgements

The authors would like to thank Zhenbin Li, Hannes Gredler, Jan Medved, Adrian Farrel, Julien Meuric and Jonathan Hardwick for the valuable discussion and comments.

8. References

8.1. Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <<http://www.rfc-editor.org/info/rfc2119>>.
- [RFC4271] Rekhter, Y., Ed., Li, T., Ed., and S. Hares, Ed., "A Border Gateway Protocol 4 (BGP-4)", RFC 4271, DOI 10.17487/RFC4271, January 2006, <<http://www.rfc-editor.org/info/rfc4271>>.
- [RFC4760] Bates, T., Chandra, R., Katz, D., and Y. Rekhter, "Multiprotocol Extensions for BGP-4", RFC 4760, DOI 10.17487/RFC4760, January 2007, <<http://www.rfc-editor.org/info/rfc4760>>.
- [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>>.
- [RFC7752] Gredler, H., Ed., Medved, J., Previdi, S., Farrel, A., and S. Ray, "North-Bound Distribution of Link-State and Traffic Engineering (TE) Information Using BGP", RFC 7752, DOI 10.17487/RFC7752, March 2016, <<http://www.rfc-editor.org/info/rfc7752>>.

8.2. Informative References

- [I-D.ietf-idr-te-pm-bgp]
Previdi, S., Wu, Q., Gredler, H., Ray, S.,
jeffrant@gmail.com, j., Filsfils, C., and L. Ginsberg,
"BGP-LS Advertisement of IGP Traffic Engineering
Performance Metric Extensions", draft-ietf-idr-te-pm-
bgp-06 (work in progress), June 2017.
- [RFC4272] Murphy, S., "BGP Security Vulnerabilities Analysis",
RFC 4272, DOI 10.17487/RFC4272, January 2006,
<<http://www.rfc-editor.org/info/rfc4272>>.
- [RFC4674] Le Roux, J., Ed., "Requirements for Path Computation
Element (PCE) Discovery", RFC 4674, DOI 10.17487/RFC4674,
October 2006, <<http://www.rfc-editor.org/info/rfc4674>>.
- [RFC5441] Vasseur, JP., Ed., Zhang, R., Bitar, N., and JL. Le Roux,
"A Backward-Recursive PCE-Based Computation (BRPC)
Procedure to Compute Shortest Constrained Inter-Domain
Traffic Engineering Label Switched Paths", RFC 5441,
DOI 10.17487/RFC5441, April 2009,
<<http://www.rfc-editor.org/info/rfc5441>>.
- [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, DOI 10.17487/RFC6952, May 2013,
<<http://www.rfc-editor.org/info/rfc6952>>.

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

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

Email: dhruv.ietf@gmail.com

Jeff Tantsura
Individual
US

Email: jefftant.ietf@gmail.com

Kenji Kumaki
KDDI Corporation
Garden Air Tower, Iidabashi, Chiyoda-ku
Tokyo 102-8460
Japan

Email: ke-kumaki@kddi.com

Tomoki Murai
Furukawa Network Solution Corp.
5-1-9, Higashi-Yawata, Hiratsuka
Kanagawa 254-0016
Japan

Email: murai@fnsc.co.jp

PCE Working Group
Internet-Draft
Intended status: Standards Track
Expires: April 11, 2016

V. Kondreddy
M. Negi
Huawei Technologies
October 9, 2015

Optimizations of PCEP Link-State(LS) Synchronization Procedures
draft-kondreddy-pce-pcep-ls-sync-optimizations-00

Abstract

For a Path Computation Element (PCE) to perform its computations, it is important that Link-State (and TE) information be complete and accurate each time. This requires a reliable Link-State Synchronization mechanism between the PCE and path computation clients (PCCs), and between cooperating PCEs. The basic mechanism for Link-State Synchronization is part of the PCEP Link-State (and TE) draft. This draft presents motivations for optimizations to the base PCEP Link-State (and TE) procedure and specifies the required Path Computation Element Communication Protocol (PCEP) extensions.

Status of This Memo

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

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

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

This Internet-Draft will expire on April 11, 2016.

Copyright Notice

Copyright (c) 2015 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. LS Synchronization Avoidance	4
3.1. Motivation	4
3.2. LS Synchronization Avoidance Procedure	4
4. Incremental LS Synchronization	8
4.1. Motivation	8
4.2. Incremental Synchronization Procedure	9
5. PCE-triggered Initial Synchronization	11
5.1. Motivation	11
5.2. PCE-triggered Initial LS Synchronization Procedure	11
6. PCE-triggered Re-synchronization	12
6.1. Motivation	12
6.2. PCE-triggered LS Re-synchronization Procedure	12
7. PCEP Extensions	13
7.1. Link-State(LS) Report Message	13
7.2. Capability Advertisement	14
7.3. Advertising Support of Synchronization Optimizations	14
8. IANA Considerations	15
8.1. PCEP-Error Object	15
8.2. PCEP TLV Type Indicators	16
8.3. LS-CAPABILITY Flags	16
9. Manageability Considerations	17
9.1. Control of Function and Policy	17
9.2. Information and Data Models	17
9.3. Liveness Detection and Monitoring	17
9.4. Verify Correct Operations	17
9.5. Requirements On Other Protocols	17
9.6. Impact On Network Operations	18
10. Security Considerations	18
11. Acknowledgement	18
12. References	18
12.1. Normative References	18
12.2. Informative References	19
Appendix A. Contributor Addresses	20
Authors' Addresses	20

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.dhodylee-pce-pcep-ls] describes a set of extensions to PCEP to provide Link-State (and TE) distribution. This draft presents motivations for optimizations to the base PCEP Link-State transport procedure and specifies the required Path Computation Element Communication Protocol (PCEP) extensions. This draft specifies following optimizations for Link-State Synchronization and the corresponding PCEP procedures and extensions:

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

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.

This document uses the following terms defined in [I-D.dhodylee-pce-pcep-ls]: LSRpt, Link-State (and TE), LS

Link-State (and TE): "LS" is interchangeably used for the keyword "Link-State (and TE)".

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

The following terms are defined in this document:

LS-DB: Link-State (and TE) Database.

LS Sync: LS Synchronization, an operation to send LS information synchronization request to PCC by LSRpt message with LS-ID 0.

LS-Info: One LS information (i.e. LS Node/Link/Prefix defined in I-D.dhodylee-pce-pcep-ls).

3. LS Synchronization Avoidance

3.1. Motivation

The purpose of LS Synchronization is to provide a checkpoint-in-time state replica of a PCC's Link-State (and TE) information in a PCE. LS Synchronization is performed immediately after the initialization phase ([RFC5440]). [I-D.dhodylee-pce-pcep-ls] describes the basic mechanism for LS synchronization.

LS Synchronization is not always necessary following a PCEP session restart. If the Link-State (and TE) information 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 PCE to become fully operational.

3.2. LS Synchronization Avoidance Procedure

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

The LS-DB Version Number, carried in LS-DB-VERSION TLV (see Section 7.2), is owned by a PCC and it MUST be incremented by 1 for each successive change in the PCC's LS-DB. The LS-DB 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 LS re-synchronization, the PCE speaker receiving this node should send back a PCErr with Error-type TBD1 Error-value 1 'Received an invalid LS-DB Version Number', and close the PCEP session. Operations that trigger a change to the local LS-DB include but not limited to -

- a change in the link status or attributes(i.e. bandwidth, metric etc), addition or deletion of link.

- a change in the node attributes, addition or deletion of node.
- a change in the prefix attributes, addition or deletion of prefix.

LS Synchronization avoidance is advertised on a PCEP session during session startup using the LS-INCLUDE-DB-VERSION (S) bit in the LS-CAPABILITY TLV (see Section 7.3). The peer may move in the network, either physically or logically, which may cause its connectivity details and transport-level identity (such as IP address) to change. To ensure that a PCEP peer can recognize a previously connected peer even in case of such mobility, each PCEP peer includes the SPEAKER-ENTITY-ID TLV in the OPEN message. SPEAKER-ENTITY-ID TLV is described in [I-D.ietf-pcep-stateful-sync-optimizations]

If both PCEP speakers set the S flag in the OPEN object's LS-CAPABILITY TLV to 1, the PCC MUST include the LS-DB-VERSION TLV in each LS object of the LSRpt message. If the LS-DB-VERSION TLV is missing in a LSRpt message, the PCE will generate an error with Error-Type 6 (mandatory object missing) and Error-Value TBD2 'LS-DB-VERSION TLV missing' and close the session. If LS Synchronization avoidance has not been enabled on a PCEP session, the PCC SHOULD NOT include the LS-DB-VERSION TLV in the LS object and the PCE SHOULD ignore it, if it were to receive one.

If a PCE's LS-DB survived the restart of a PCEP session, the PCE will include the LS-DB-VERSION TLV in its OPEN object, and the TLV will contain the last LS-DB Version Number received on an LS Report from the PCC in the previous PCEP session. If a PCC's LS-DB survived the restart of a PCEP session, the PCC will include the LS-DB-VERSION TLV in its OPEN object and the TLV will contain the latest LS-DB Version Number. If a PCEP speaker's LS-DB did not survive the restart of a PCEP session, the PCEP speaker MUST NOT include the LS-DB-VERSION TLV in the OPEN object.

If both PCEP speakers include the LS-DB-VERSION TLV in the OPEN Object and the TLV values match, the PCC MAY skip LS Synchronization. Otherwise, the PCC MUST perform complete LS Synchronization. If the PCC attempts to skip LS Synchronization (i.e., the SYNC Flag = 0 on the first LS Report from the PCC, the PCE MUST send back a PCerr with Error-Type TBD1 Error-Value 2 'LS-DB version mismatch', and close the PCEP session.

If LS Synchronization is required, then prior to completing the initialization phase, the PCE MUST mark any LS-Infos in the LS-DB that were previously reported by the PCC as stale. When the PCC reports a LS-Info during LS Synchronization, if the LS-Info already exists in the LS-DB, the PCE MUST update the LS-DB and clear the stale marker from the LS-Info. When it has finished LS

Synchronization, the PCC MUST immediately send an end of LS Synchronization marker. The end of synchronization marker is a LS Report (LSRpt) message with an LS object containing a LS-ID of 0 and with the SYNC flag set to 0. The LS-DB-VERSION TLV MUST be included in this LSRpt message. On receiving this LS Report, the PCE MUST purge any LS-Infos from the LS-DB that are still marked as stale. It should be noted that PCE may receive the same Link-state and TE information from multiple PCCs and the purging should take that into account.

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

Since a PCE does not make changes to the LS-DB 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 LS-DB-VERSION TLV, were it to receive one from a PCE.

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

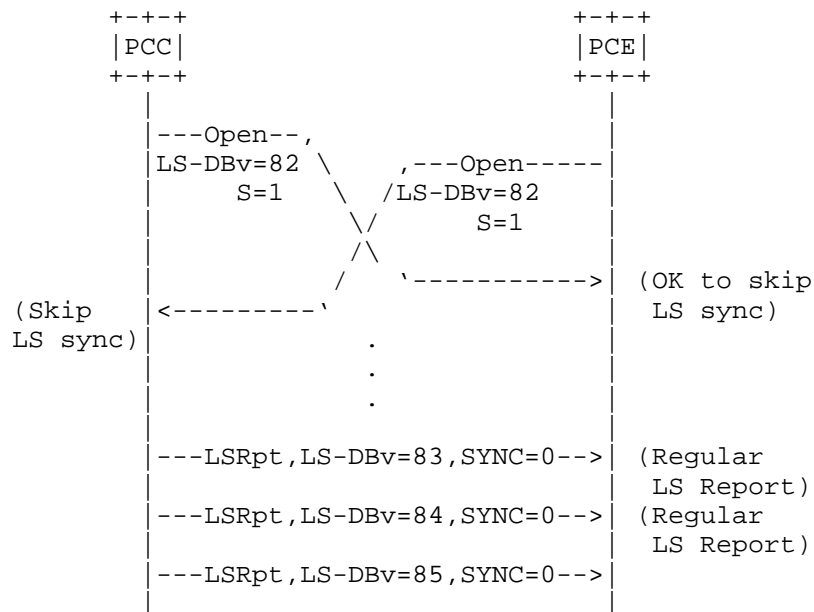


Figure 1: LS Synchronization Skipped

Figure 2 shows an example sequence where the LS Synchronization is performed due to LS-DB Version mismatch during the PCEP session setup. Note that the same LS Synchronization sequence would happen if either the PCC or the PCE would not include the LS-DB-VERSION TLV in their respective Open messages.

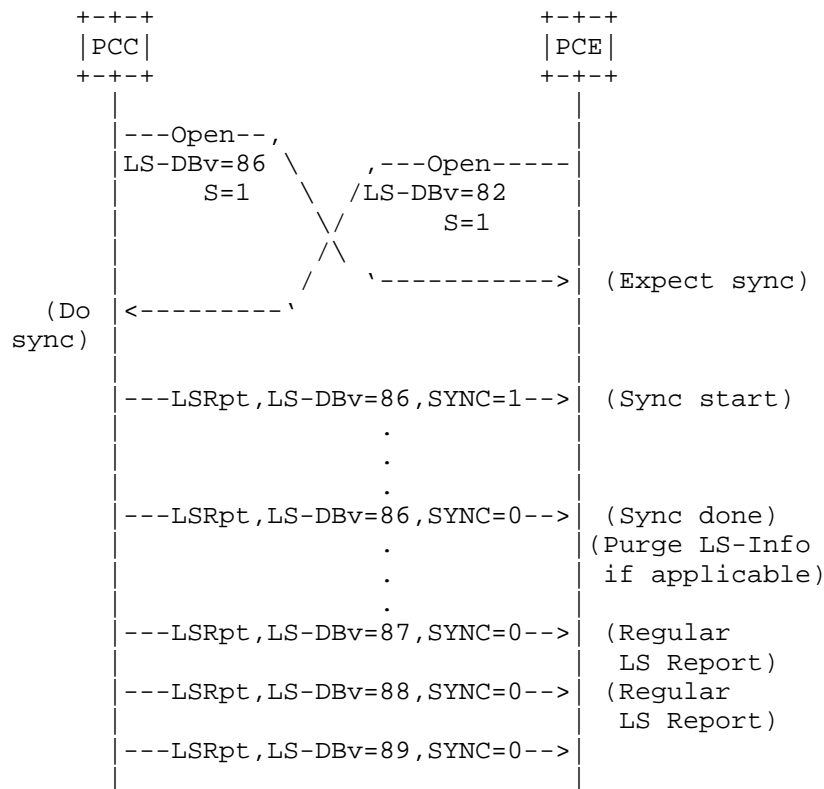


Figure 2: LS Synchronization Performed

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

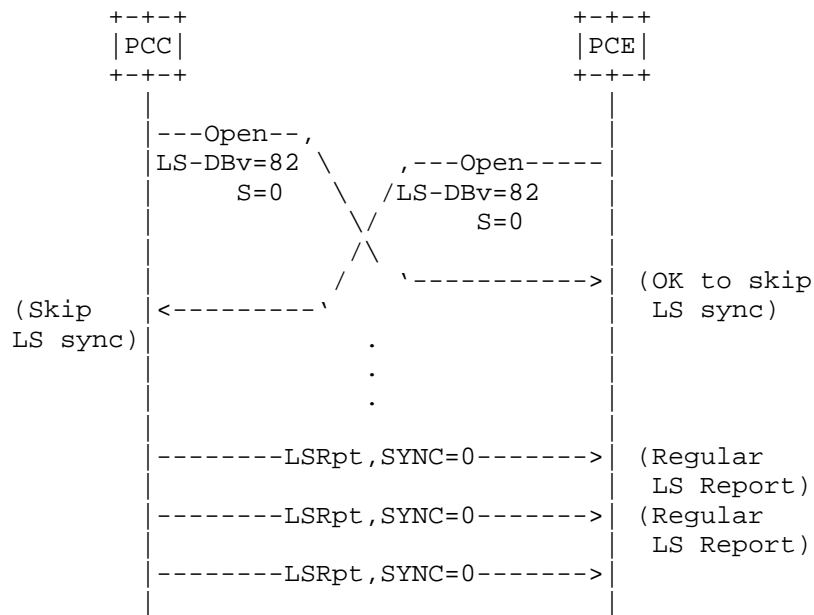


Figure 3: LS Synchronization Skipped, no LS-DB-VERSION TLVs sent from PCC

4. Incremental LS Synchronization

[I-D.dhodylee-pce-pcep-ls] describes the LS synchronization mechanism during session initialization between PCCs and PCEs. During the LS synchronization, a PCC sends the information of its LS-DB to the PCE based on the local policy. In order to reduce the LS Synchronization overhead when there is a small number of LS-DB change in the network between PCEP session restart, this section defines a mechanism for incremental (Delta) LS synchronization.

4.1. Motivation

According to [I-D.dhodylee-pce-pcep-ls], if a PCEP session restarts, PCCs send snapshot of LS-DB information to the PCE, though LS-DB did not change. And as per Section 3 (LS Synchronization Avoidance Procedure), if there is a change in a small number of LS-Infos. PCC yet sends complete snapshot of LS-DB information to the PCE, which takes a long time and consume large communication channel bandwidth.

4.2. Incremental Synchronization Procedure

Section 3 describes LS Synchronization avoidance by using LS-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 LS-DB-VERSION TLV in the OPEN object and the LS-DB-VERSION TLV values match, the PCC MAY skip LS Synchronization. Otherwise, the PCC MUST perform LS Synchronization. Incremental LS Synchronization capability is advertised on a PCEP session during session startup using the LS-DELTA-SYNC-CAPABILITY (D) bit in the capabilities TLV (see Section 7.3). Instead of dumping full LS-DB to the PCE again, PCC synchronizes the delta (changes) as described in Figure 4 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 complete LS Synchronization procedure as described in [I-D.dhodylee-pce-pcep-ls]. If a PCC has to force complete LS Synchronization due to reasons including but not limited: (1) local policy configured at the PCC; (2) no sufficient LS-DB caches for incremental update, the PCC can set the D flag to 0. Note a PCC may have to bring down the current session and force a complete LS Synchronization with D flag set to 0 in the subsequent open message.

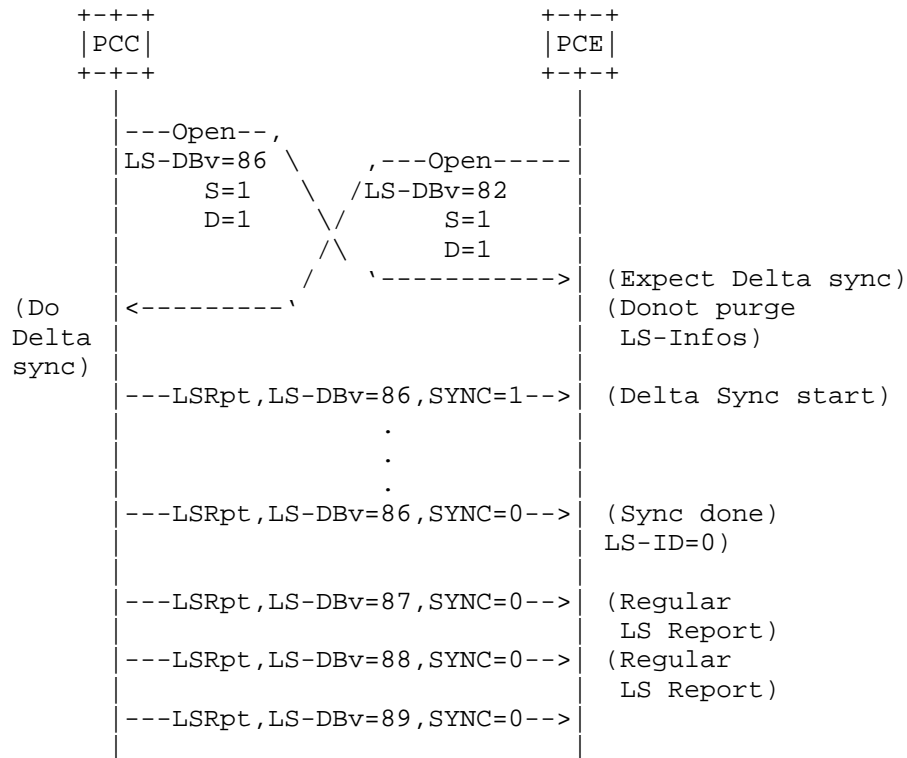


Figure 4: Incremental Synchronization Procedure

As per Section 3, the LS-DB Version Number is incremented each time a change is made to the PCC's local LS-DB. Each LS-Info is associated with the DB version at the time of change. This is needed to determine which LS-Info needs to be synchronized in incremental LS Synchronization.

PCC MAY store a then history of LS-DB change that happened between the PCEP session(s) restart in order to carry out incremental LS Synchronization. After the synchronization procedure finishes, the PCC can dump this history information. In the example shown in Figure 4, the PCC needs to store the LS-DB changes that happened between DB Version 83 to 86 and synchronizes these changes only when performing incremental LS-DB update. So a PCC needs to remember the LS-DB changes that happened when an existing PCEP session to a PCE goes down in the hope of doing incremental synchronization when the session is re-established.

If a PCC finds out it does not have sufficient information to complete incremental synchronization after advertising incremental LS Synchronization capability, it MUST send a PCErr with Error-Type TBD1 and Error-Value 3 'A PCC indicates to a PCE that it can not complete the LS synchronization' and terminate the session.

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 PCE connected to the network via one network node, it is desirable to control the timing of PCC LS 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 PCE and the network. The method proposed, i.e., allowing PCE to trigger the LS 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 LS Synchronization Procedure

Support of PCE-triggered LS Synchronization is advertised during session startup using the LS-TRIGGERED-INITIAL-SYNC (F) bit in the LS-CAPABILITY TLV (see Section 7.3).

As per [I-D.dhodylee-pce-pcep-ls], LSRpt is sent from PCC to PCE, this document extends the usage of LSRpt to trigger synchronization. Where a PCC can send a LSRpt (for LS Sync) with an LS object containing a LS-ID of 0 and with the SYNC flag set to 1. This LSRpt message is the trigger for the PCC to enter the synchronization phase and start sending LSRpt messages.

If the LS-TRIGGERED-INITIAL-SYNC capability is not advertised and the PCC receives a LSRpt with the SYNC flag set to 1, it MUST send a PCErr for LSRpt(LS Sync from PCE) with Error-Type TBD1 and Error-Value 4 'Attempt to trigger synchronization when the PCE triggered synchronization capability has not been advertised'.

A PCE MAY choose to control the LS Synchronization process. To allow PCE to do so, PCEP speakers MUST set T bit to 1 to indicate this (as described in Section 7.3). If the LS-DB version is mis-matched, it can send a LSRpt message with LS-ID = 0 and SYNC = 1 in order to trigger the LS Synchronization process. In this way, the PCE can control the sequence of LS 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. The PCC SHOULD NOT send LSRpt messages to the PCE before it triggers the LS Synchronization. This PCE-triggering capability can be applied to both full and incremental LS Synchronization. If applied to the later, the PCCs only send information that PCE does not possess, which is inferred from the LS-DB version information exchanged in the OPEN message (see Section 3.2) for detailed procedure).

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 LS-DB. Therefore, it can be beneficial to be able to re-synchronize LS-DB even after the session has been established. The PCE may use this approach to continuously sanity check its LS-DB against the network, or to recover from error conditions without having to tear down sessions.

6.2. PCE-triggered LS Re-synchronization Procedure

Support of PCE-triggered LS Synchronization is advertised during session startup using the LS-TRIGGERED-RESYNC (T) bit in the LS-CAPABILITY TLV (see Section 7.3).

The PCE triggers re-synchronization of the entire LS-DB. The PCE MUST first mark all LS-Infos in the LS-DB that were previously reported by the PCC as stale and then send a LSRpt (for LS Sync) with an LS object containing a LS-ID of 0 and with the SYNC flag set to 1. This LSRpt message is the trigger for the PCC to enter the synchronization phase and start sending LSRpt messages. After the receipt of the end-of-synchronization marker, the PCE will purge LS-Infos which were not refreshed.

If the LS-TRIGGERED-RESYNC capability is not advertised and the PCC receives a LSRpt with the SYNC flag set to 1, it MUST send a PCErr with Error-Type TBD1 and Error-Value 4 'Attempt to trigger synchronization when the TRIGGERED-SYNC capability has not been advertised'.

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

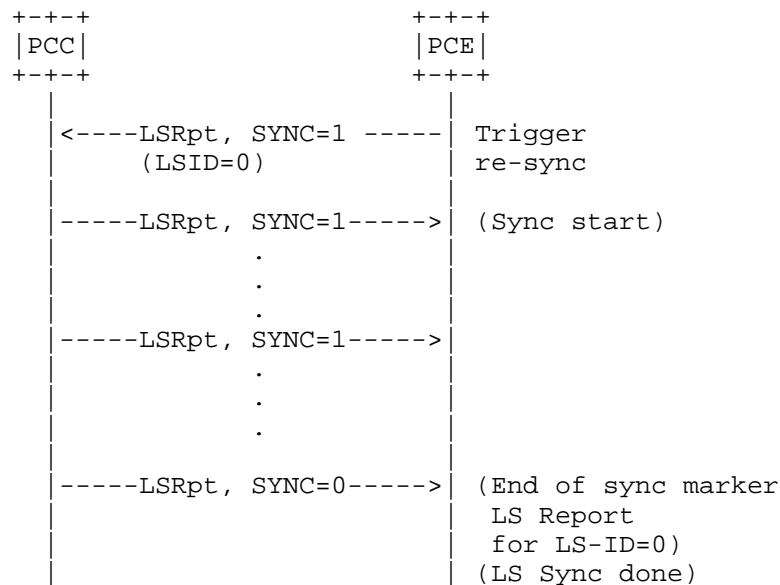


Figure 5: PCE Triggered Complete LS re-synchronization

7. PCEP Extensions

7.1. Link-State(LS) Report Message

A PCEP LS Report message (also referred to as LSRpt message) is a PCEP message sent by a PCC to a PCE to report the LS information. The definition of the LSRpt message from [I-D.dhodylee-pce-pcep-ls] is extended to use LSRpt message with LS-ID = 0 to request LS Synchronization from PCE to PCC.

If a PCC that does not support extension defined in this document receives a LSRpt message, it will act according to existing behavior of receiving invalid message. If a PCC supports the extension, but did not set the flag T or F, and receives the LSRpt message, it sends PCErr message as described earlier in section [x] and [y]. If a PCC supports the extension and set the flag T or F, and receives the LSRpt message without LS-ID as 0 and SYNC flag set, PCC will send an error message with Error-Type TBD1 Error-Value 6 'Invalid LSRpt message'.

7.2. Capability Advertisement

The LS-DB Version Number is carried in optional LS-DB-VERSION TLV that MAY be included in the OPEN object and the LS object. This TLV MUST NOT be included if LS-INCLUDE-DB-VERSION bit in LS-CAPABILITY TLV is not set.

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

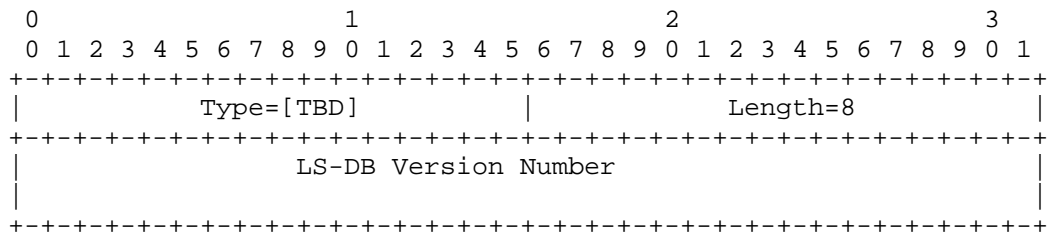


Figure 6: LS-DB-VERSION TLV format

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

7.3. Advertising Support of Synchronization Optimizations

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

New flags are defined for the LS-CAPABILITY TLV defined in [I-D.dhodylee-pce-pcep-ls].

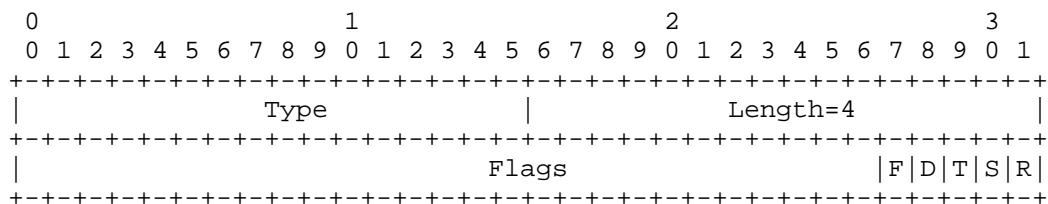


Figure 7: LS-CAPABILITY TLV Format

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

R (LS-REMOTE-BIT - 1 bit): Defined in [I-D.dhodylee-pce-pcep-ls]

S (LS-INCLUDE-DB-VERSION - 1 bit): If set to 1 by both PCEP speakers, the PCC will include the LS-DB-VERSION TLV in each LS object. See Section 3 for details.

T (LS-TRIGGERED-RESYNC - 1 bit): If set to 1 by both PCEP speakers, the PCE can trigger re-synchronization of LS-Infos at any point in the life of the session. See Section 6 for details.

D (LS-DELTA-SYNC-CAPABILITY - 1 bit): If set to 1 by a PCEP speaker, it indicates that the PCEP speaker allows incremental (delta) state synchronization. See Section 4 for details.

F (LS-TRIGGERED-INITIAL-SYNC - 1 bit): If set to 1 by both PCEP speakers, the PCE SHOULD trigger initial (first) LS synchronization. See Section 5 for details.

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= TBD2	[RFC5440] This document
TBD1	LS-DB-VERSION TLV missing LS synchronization error Error-Value= TBD(suggested value 1):Received an invalid LSDB Version Number	This document This document
	Error-Value= TBD(suggested value 2): LS-DB version mismatch.	This document
	Error-Value=TBD(suggested value 3): PCC indicates to a PCE that it cannot complete the LS Synchronization	This document
	Error-Value=TBD(suggested value 4): Attempt to trigger a synchronization when the PCE triggered synchronization capability has not been advertised.	This document
	Error-Value=TBD(suggested value 5): LS-DB-VERSION TLV Missing when LS synchronization avoidance is enabled.	This document
	Error-Value=TBD(suggested value 6): Invalid LSRpt message.	This document

8.2. PCEP TLV Type Indicators

This document defines the following new PCEP TLVs:

Value	Meaning	Reference
TBD	LS-DB-VERSION	This document

8.3. LS-CAPABILITY Flags

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

Bit	Description	Reference
TBD		
(Suggested value 30)	LS-INCLUDE-DB-VERSION	This document
(Suggested value 29)	LS-TRIGGERED-RESYNC	This document
(Suggested value 28)	LS-DELTA-SYNC-CAPABILITY	This document
(Suggested value 27)	LS-TRIGGERED-INITIAL-SYNC	This document

9. Manageability Considerations

All manageability requirements and considerations listed in [RFC5440] and [I-D.dhodylee-pce-pcep-ls] 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.

9.2. Information and Data Models

PCEP session configuration and information in the PCEP MIB module SHOULD be extended to include advertised LS Capabilities, LS-DB Version Number and LS synchronization status, Message statistics.

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 section 8.4 of [RFC5440] also apply to PCEP protocol extensions defined in this document. In addition to monitoring parameters defined in [RFC5440], a PCEP implementation with LS-DB SHOULD provide the following parameters:

- o Total number of LSRpt(Synchronization request) requests
- o LS-DB Version Number and synchronization status

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 section 8.6 of [RFC5440] also apply to PCEP protocol extensions defined in this document.

Additionally, a PCEP implementation SHOULD allow a limit to be placed on the amount and rate of LSRpt messages sent by a PCEP speaker and processed by the peer. It SHOULD also allow sending a notification when a rate threshold is reached.

10. Security Considerations

The security considerations listed in [I-D.dhodylee-pce-pcep-ls] apply to this document as well. However, because the protocol modifications outlined in this document allow the PCE to control LS Re-synchronization timing and sequence, it also introduces a new attack vector: 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 is free to drop any trigger re-synchronization request without additional processing.

11. Acknowledgement

The document borrows some of the text and structure from [I-D.ietf-pce-stateful-sync-optimizations].

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, <<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>>.
- [I-D.dhodylee-pce-pcep-ls] Dhody, D., Lee, Y., and D. Ceccarelli, "PCEP Extension for Distribution of Link-State and TE Information.", draft-dhodylee-pce-pcep-ls-00 (work in progress), September 2015.

12.2. Informative References

[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-03 (work in
progress), October 2015.

Appendix A. Contributor Addresses

Dhruv Dhody
Huawei Technologies
Divyashree Techno Park, Whitefield
Bangalore, Karnataka 560037
India

EMail: dhruv.ietf@gmail.com

Authors' Addresses

Venugopal Reddy Kondreddy
Huawei Technologies
Divyashree Techno Park, Whitefield
Bangalore, Karnataka 560037
India

EMail: venugopalreddyk@huawei.com

Mahendra Singh Negi
Huawei Technologies
Divyashree Techno Park, Whitefield
Bangalore, Karnataka 560037
India

EMail: mahendrasingh@huawei.com

PCE Working Group
Internet-Draft
Intended status: Standards Track
Expires: January 7, 2016

I. Minei
Google, Inc.
E. Crabbe

S. Sivabalan
Cisco Systems, Inc.
H. Ananthakrishnan
Packet Design
X. Zhang
Huawei Technologies
Y. Tanaka
NTT Communications Corporation
July 6, 2015

PCEP Extensions for Establishing Relationships Between Sets of LSPs
draft-minei-pce-association-group-02

Abstract

This document introduces a generic mechanism to create a grouping of LSPs in the context of a 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 a stateful PCE as well as a stateless 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 January 7, 2016.

Copyright Notice

Copyright (c) 2015 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. ASSOCIATION Object	4
4.1. Object Definition	4
4.2. Object Encoding in PCEP messages	5
4.3. Processing Rules	8
5. IANA Considerations	9
6. Security Considerations	9
7. Acknowledgements	9
8. Contributors	10
9. References	10
9.1. Normative References	10
9.2. Informative References	10
Authors' Addresses	10

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) as well as Generalized MPLS (GMPLS) 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 a stateful PCE and a stateless 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 re-optimized 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.

For a stateless PCE, it might be useful to associate a path computation request to an association group, thus enabling it to associate a common set of configuration parameters or behaviors with the request.

Rather than creating separate mechanisms for each use case, this draft defines a generic mechanism that can be reused as needed.

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 can be applied to LSPs originating at the same head end or different head ends. For LSPs originating at the same head end, the association can be initiated by either the PCC (head end) or by a PCE. Only a stateful PCE can initiate an association for LSPs originating at different head ends. For both

cases, the association is uniquely identified by the combination of an association identifier and the address of the PCE peer that created the association.

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. An LSP may join multiple association groups, of different or of the same type.

In the case of a stateless PCE, associations are created out of band, and PCEP peers should be aware of the association and its significance outside of the protocol.

4. ASSOCIATION Object

4.1. Object Definition

Creation of an association group and modifications to its membership can be initiated by either the PCE or the PCC. Association groups and their memberships are defined using the ASSOCIATION object for stateful PCE.

ASSOCIATION Object-Class is to be assigned by IANA (TBD).

ASSOCIATION Object-Type is 1 for IPv4 and its format is shown in Figure 1:

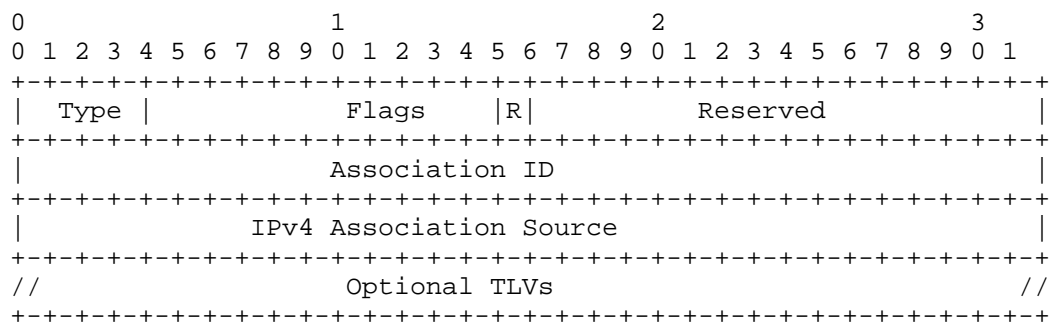


Figure 1: The IPv4 ASSOCIATION Object format

ASSOCIATION Object-Type is 2 for IPv6 and its format is shown in Figure 2:

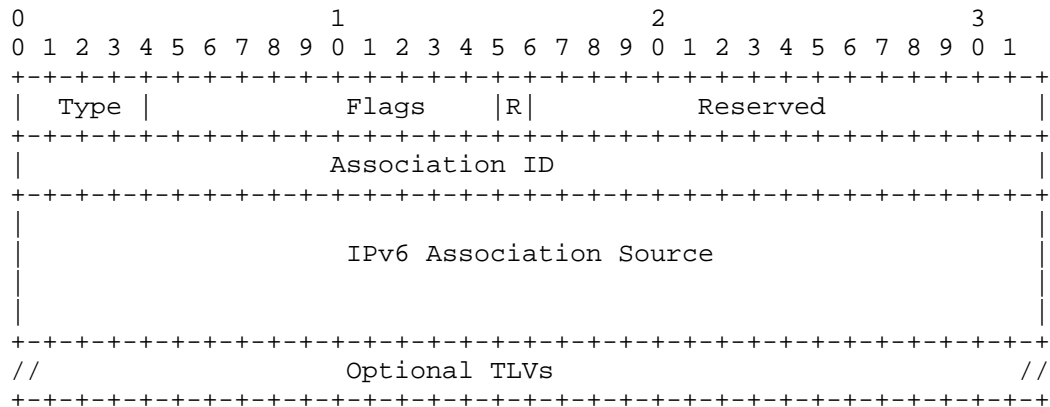


Figure 2: The IPv6 ASSOCIATION Object format

Type: 4 bits - the association type (for example protection). The association type will be defined in separate documents.

Flags: 12 bits - The following flags are currently defined:

R (Removal - 1 bit): when set, the requesting PCE peer requires the removal of an LSP from the association group.

Reserved: MUST be set to 0 and ignored upon receipt.

Association ID: 32 bits - the identifier of the association group. When combined with Type and Association Source, this value uniquely identifies an association group. The value 0xffffffff and 0x0 are reserved. The value 0xffffffff is used to indicate all association groups.

Association Source: 4 or 16 bytes - An IPv4 or IPv6 address, which is associated to the PCE peer that originated the association.

Optional TLVs: Variable - no TLVs are defined in this document.

4.2. Object Encoding in PCEP messages

The ASSOCIATION Object is OPTIONAL and MAY be carried in the Path Computation Update (PCUpd), Path Computation Report (PCRpt) and Path Computation Initiate (PCinit) messages.

When carried in PCRpt message, it is used to report the association group membership information pertaining to a LSP to a stateful PCE. It can also be used to remove an LSP from one or more association

groups by setting the R flag to 1. Unless, a PCE wants to delete an association from an LSP, it does not need to carry the ASSOCIATION object while updating other LSP attributes using the PCUpd message.

The PCRpt message is defined in [I-D.ietf-pce-stateful-pce] and updated as below:

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

Where:

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

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

Where:

```
<association-list> ::= <ASSOCIATION> [<association-list>]
```

When an LSP is delegated to a stateful PCE, the stateful PCE can initiate a new association group for this LSP, or associate it with one or more existing association groups. This is done by including the ASSOCIATION Object in a PCUpd message or in a PCInit message. A stateful PCE can also remove a delegated LSP from one or more association groups by setting the R flag to 1.

The PCUpd message is defined in [I-D.ietf-pce-stateful-pce] and updated as below:

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

Where:

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

```
<update-request> ::= <SRP>
                    <LSP>
                    [<association-list>]
                    <path>
```

Where: <association-list> ::= <ASSOCIATION> [<association-list>]

The PCInitiate message is defined in [I-D.ietf-pce-pce-initiated-lsp] and updated as below:

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

Where:

```
<PCE-initiated-lsp-list> ::=  
<PCE-initiated-lsp-request>[<PCE-initiated-lsp-list>]  
  
<PCE-initiated-lsp-request> ::=  
(<PCE-initiated-lsp-instantiation> | <PCE-initiated-lsp-deletion>)  
  
<PCE-initiated-lsp-instantiation> ::= <SRP>  
                                       <LSP>  
                                       <END-POINTS>  
                                       <ERO>  
                                       [<association-list>]  
                                       [<attribute-list>]
```

Where:

```
<association-list> ::= <ASSOCIATION> [<association-list>]
```

In case of passive stateful or stateless PCE, the ASSOCIATION Object is OPTIONAL and MAY be carried in the Path Computation Request (PCReq) message.

When carried in a PCReq message, the ASSOCIATION Object is used to associate the path computation request to an association group, the association might be further informed via PCRpt message in case of passive stateful PCE later or it might be created out of band in case of stateless PCE.

The PCReq message is defined in [RFC5440] and updated in [I-D.ietf-pce-stateful-pce], it is further updated below for association:

```

<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>]
              [<association-list>]
              [<RRO> [<BANDWIDTH>]]
              [<IRO>]
              [<LOAD-BALANCING>]

```

Where:

```

<association-list> ::= <ASSOCIATION> [<association-list>]

```

Note that LSP object MAY be present for the passive stateful PCE.

4.3. Processing Rules

Both a PCC and a PCE can create one or more association groups for an LSP. But a PCE peer cannot add new members for association group created by another peer. If a PCC receives a PCUpd or a PCInitiate message including an ASSOCIATION Object but the sender address does not match the association source, a PCErr message MUST be sent with Error-Type = TBD2 (Association Error) and Error-value= 1 (association source and sender source mismatch in PCUpd). Error handling for situations such as PCE failures after association groups are created and other scenarios will be included in future versions of this draft.

If a PCE peer does not recognize the ASSOCIATION object, it MUST return a PCErr message with Error-Type "Unknown Object" as described in [RFC5440]. If a PCE peer is unwilling or unable to process the ASSOCIATION object, it MUST return a PCErr message with the Error-Type "Not supported object" and follow the relevant procedures described in [RFC5440].

5. IANA Considerations

The "PCEP Parameters" registry contains a subregistry "PCEP Objects". This document request IANA to allocate the values from this registry.

Object-Class Value	Name	Reference
TBD	Association Object-Type 1: IPv4 2: IPv6	This document

This document requests IANA to create a subregistry of the "PCEP Parameters" for the bits carried in the Flags field of the ASSOCIATION object. The subregistry is called "ASSOCIATION Flags Field".

The field contains 12 bits numbered from bit 0 as the most significant bit.

Bit;	Name;	Description	Reference
15	R:	Removal	This document

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

Error-Type	Meaning	Reference
TBD	Error-Value=1: association source and sender source does not match	This document

6. 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.

7. Acknowledgements

We would like to thank Yuji Kamite and Joshua George for their contributions to this document. Also Thank Venugopal Reddy and Cyril Margaria for their useful comments.

8. Contributors

Dhruv Dhody
Huawei Technologies
Divyashree Techno Park, Whitefield
Bangalore, Karnataka 560037
India
Email: dhruv.ietf@gmail.com

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-04 (work in progress), April 2015.
- [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-11 (work in progress), April 2015.
- [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.

9.2. Informative References

- [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

Email: edward.crabbe@gmail.com

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

Email: msiva@cisco.com

Hariharan Ananthakrishnan
Packet Design

Email: hari@packetdesign.com

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

Email: zhang.xian@huawei.com

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: May 12, 2016

I. Minei
Google, Inc.
E. Crabbe

S. Sivabalan
Cisco Systems, Inc.
H. Ananthakrishnan
Packet Design
X. Zhang
Huawei Technologies
Y. Tanaka
NTT Communications Corporation
November 9, 2015

PCEP Extensions for Establishing Relationships Between Sets of LSPs
draft-minei-pce-association-group-04

Abstract

This document introduces a generic mechanism to create a grouping of LSPs in the context of a 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 a stateful PCE as well as a stateless 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 May 12, 2016.

Copyright Notice

Copyright (c) 2015 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	4
4. ASSOCIATION Object	4
4.1. Object Definition	4
4.1.1. Global Association Source TLV	6
4.1.2. Extended Association ID TLV	6
4.2. Object Encoding in PCEP messages	7
4.3. Processing Rules	9
5. IANA Considerations	10
6. Security Considerations	11
7. Acknowledgements	11
8. Contributors	11
9. References	11
9.1. Normative References	11
9.2. Informative References	12
Authors' Addresses	12

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) as well as Generalized MPLS (GMPLS) 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 a stateful PCE and a stateless PCE.

2. Terminology

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

The following term is defined in this document:

Association Timeout Interval: when a PCEP session is terminated, a PCC waits for this time period before deleting associations created by the 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 re-optimized 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.

For a stateless PCE, it might be useful to associate a path computation request to an association group, thus enabling it to associate a common set of configuration parameters or behaviors with the request.

Rather than creating separate mechanisms for each use case, this draft defines a generic mechanism that can be reused as needed.

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 can be applied to LSPs originating at the same head end or different head ends. For LSPs originating at the same head end, the association can be initiated by either the PCC (head end) or by a PCE. Only a stateful PCE can initiate an association for LSPs originating at different head ends. For both cases, the association is uniquely identified by the combination of an association identifier and the address of the node that created the association.

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. An LSP may join multiple association groups, of different or of the same type.

In the case of a stateless PCE, associations are created out of band, and PCEP peers should be aware of the association and its significance outside of the protocol.

Association groups can be created by both PCC and PCE. When a PCC's PCEP session with a PCE terminates unexpectedly, the PCC cleans up associations (as per the processing rules in this document).

4. ASSOCIATION Object

4.1. Object Definition

Creation of an association group and modifications to its membership can be initiated by either the PCE or the PCC. Association groups and their memberships are defined using the ASSOCIATION object for stateful PCE.

ASSOCIATION Object-Class is to be assigned by IANA (TBD).

ASSOCIATION Object-Type is 1 for IPv4 and its format is shown in Figure 1:

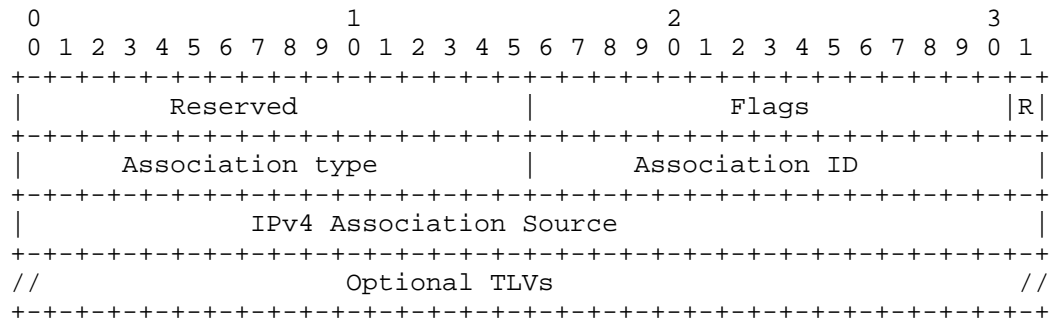


Figure 1: The IPv4 ASSOCIATION Object format

ASSOCIATION Object-Type is 2 for IPv6 and its format is shown in Figure 2:

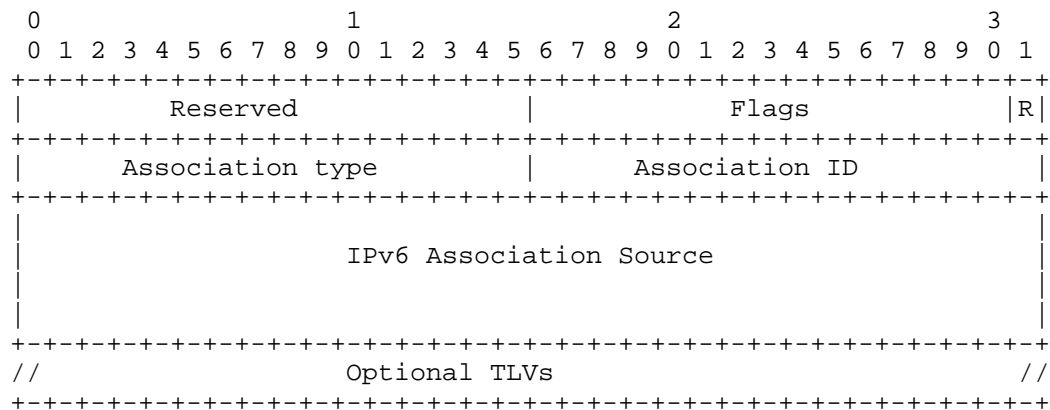


Figure 2: The IPv6 ASSOCIATION Object format

Reserved: 16 bits - MUST be set to 0 and ignored upon receipt.

Flags: 16 bits - The following flags are currently defined:

R (Removal - 1 bit): when set, the requesting PCE peer requires the removal of an LSP from the association group.

Association type: 16 bits - the association type (for example protection). The association type will be defined in separate documents.

Association ID: 16 bits - the identifier of the association group. When combined with Type and Association Source, this value uniquely identifies an association group. The value 0xffff and 0x0 are reserved. The value 0xffff is used to indicate all association groups.

Association Source: 4 or 16 bytes - An IPv4 or IPv6 address, which is associated to the node that originated the association.

Optional TLVs: The optional TLVs follow the PCEP TLV format of [RFC5440]. This document defines two optional TLVs.

4.1.1. Global Association Source TLV

The Global Association Source TLV is an optional TLV for use in the Association Object.

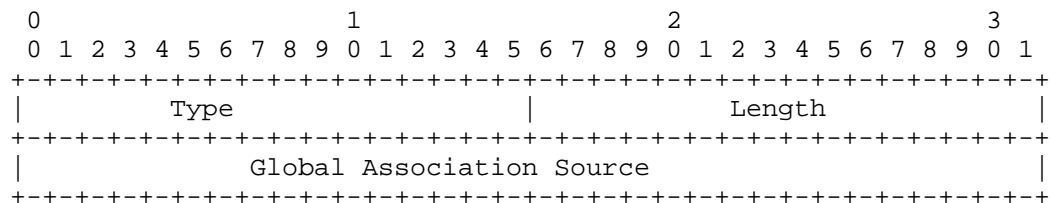


Figure 3: The Global Association Source TLV format

Type: To be allocated by IANA

Length: Fixed value of 4 bytes

Global Association Source: as defined in [RFC6780]

4.1.2. Extended Association ID TLV

The Extended Association ID TLV is an optional TLV for use in the Association Object.

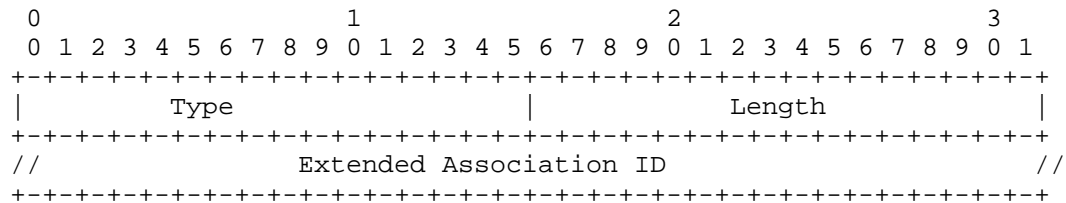


Figure 4: The Extended Association ID TLV format

Type: To be allocated by IANA

Length: variable

Extended Association ID: as defined in [RFC6780]

4.2. Object Encoding in PCEP messages

The ASSOCIATION Object is OPTIONAL and MAY be carried in the Path Computation Update (PCUpd), Path Computation Report (PCRpt) and Path Computation Initiate (PCinit) messages.

When carried in PCRpt message, it is used to report the association group membership information pertaining to a LSP to a stateful PCE. It can also be used to remove an LSP from one or more association groups by setting the R flag to 1. Unless, a PCE wants to delete an association from an LSP, it does not need to carry the ASSOCIATION object while updating other LSP attributes using the PCUpd message.

The PCRpt message is defined in [I-D.ietf-pce-stateful-pce] and updated as below:

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

Where:

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

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

Where:

```
<association-list> ::= <ASSOCIATION> [<association-list>]
```


When an LSP is delegated to a stateful PCE, the stateful PCE can initiate a new association group for this LSP, or associate it with one or more existing association groups. This is done by including the ASSOCIATION Object in a PCUpd message or in a PCInit message. A stateful PCE can also remove a delegated LSP from one or more association groups by setting the R flag to 1.

The PCUpd message is defined in [I-D.ietf-pce-stateful-pce] and updated as below:

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

Where:

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

<update-request> ::= <SRP>
                    <LSP>
                    [<association-list>]
                    <path>
```

Where: <association-list> ::= <ASSOCIATION> [<association-list>]

The PCInitiate message is defined in [I-D.ietf-pce-pce-initiated-lsp] and updated as below:

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

Where:

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

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

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

Where:

```
<association-list> ::= <ASSOCIATION> [<association-list>]
```

In case of passive stateful or stateless PCE, the ASSOCIATION Object is OPTIONAL and MAY be carried in the Path Computation Request (PCReq) message.

When carried in a PCReq message, the ASSOCIATION Object is used to associate the path computation request to an association group, the association might be further informed via PCRpt message in case of passive stateful PCE later or it might be created out of band in case of stateless PCE.

The PCReq message is defined in [RFC5440] and updated in [I-D.ietf-pce-stateful-pce], it is further updated below for association:

```
<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>]
              [<association-list>]
              [<RRO> [<BANDWIDTH>]]
              [<IRO>]
              [<LOAD-BALANCING>]
```

Where:

```
<association-list> ::= <ASSOCIATION> [<association-list>]
```

Note that LSP object MAY be present for the passive stateful PCE.

4.3. Processing Rules

Both a PCC and a PCE can create one or more association groups for an LSP. But a PCE peer cannot add new members for association group created by another peer. If a PCE peer does not recognize the ASSOCIATION object, it MUST return a PCErr message with Error-Type "Unknown Object" as described in [RFC5440]. If a PCE peer is unwilling or unable to process the ASSOCIATION object, it MUST return a PCErr message with the Error-Type "Not supported object" and follow the relevant procedures described in [RFC5440].

The association timeout interval is as a PCC-local value that can be operator-configured or computed by the PCC based on local policy and is used in the context of cleaning up associations on session failure. The association timeout must be set to a value no larger

than the state timeout interval (defined in [I-D.ietf-pce-stateful-pce]) and larger than the delegation timeout interval (defined in [I-D.ietf-pce-stateful-pce]).

When a PCC's PCEP session with the PCE terminates unexpectedly, the PCC MUST wait for the association timeout interval before cleaning up the association. If this PCEP session can be re-established before the association timeout interval time expires, no action is taken to clean the association created by this PCE. During the time window of the redelegation timeout interval and the association timeout interval, the PCE, after re-establishing the session, can also ask for redelegation following the procedure defined in [I-D.ietf-pce-stateful-pce] and [I-D.ietf-pce-pce-initiated-lsp]. When the association timeout interval timers expires, the PCC clears all the associations which are not delegated to any PCEs.

Upon LSP delegation revocation, the PCC MAY clear the association created by the related PCE, but in order to avoid traffic loss, it can perform this in a make-before-break fashion, which is the same as what is defined in Stateful pce [I-D.ietf-pce-stateful-pce] for handling LSP state cleanup.

Error handling for situations for multiple PCE scenarios will be included in future versions of this draft.

5. IANA Considerations

The "PCEP Parameters" registry contains a subregistry "PCEP Objects". This document request IANA to allocate the values from this registry.

Object-Class Value	Name	Reference
TBD	Association Object-Type 1: IPv4 2: IPv6	This document

This document defines the following new PCEP TLVs:

Value	Meaning	Reference
TBD	Global Association Source	This document
TBD	Extended Association Id	This document

This document requests IANA to create a subregistry of the "PCEP Parameters" for the bits carried in the Flags field of the ASSOCIATION object. The subregistry is called "ASSOCIATION Flags Field".

The field contains 12 bits numbered from bit 0 as the most significant bit.

Bit	Name	Description	Reference
15	R	Removal	This document

6. 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.

7. Acknowledgements

We would like to thank Yuji Kamite and Joshua George for their contributions to this document. Also Thank Venugopal Reddy and Cyril Margaria for their useful comments.

8. Contributors

Dhruv Dhody
Huawei Technologies
Divyashree Techno Park, Whitefield
Bangalore, Karnataka 560037
India
Email: dhruv.ietf@gmail.com

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-05 (work in progress), October 2015.
- [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-12 (work in progress), October 2015.

- [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>>.
- [RFC6780] Berger, L., Le Faucheur, F., and A. Narayanan, "RSVP ASSOCIATION Object Extensions", RFC 6780, DOI 10.17487/RFC6780, October 2012, <<http://www.rfc-editor.org/info/rfc6780>>.

9.2. 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, <<http://www.rfc-editor.org/info/rfc4657>>.

Authors' Addresses

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

Email: inaminei@google.com

Edward Crabbe

Email: edward.crabbe@gmail.com

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

Email: msiva@cisco.com

Hariharan Ananthakrishnan
Packet Design

Email: hari@packetdesign.com

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

Email: zhang.xian@huawei.com

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: April 8, 2016

S. Sivabalan
C. Filsfils
Cisco Systems, Inc.
J. Tantsura
Ericsson
J. Hardwick
Metaswitch Networks
October 6, 2015

Conveying policies associated with traffic engineering paths over PCEP
session
draft-sivabalan-pce-policy-identifier-00.txt

Abstract

This document describes a simple extension to the Path Computation Element (PCE) Communication Protocol (PCEP) using which a PCEP speaker can enforce one or more policies on the other PCEP speaker. A policy is represented by a numeric value which can be interpreted only by the receiving PCEP speaker. Using the proposed extension, a path computation client (PCC) can signal one or more policies that must be taken into consideration by a PCE during path computation. Similarly, when initiating or updating a path, a stateful PCE can signal one or more policies (e.g., traffic steering rules) that a PCC is expected to apply to the path.

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, 2016.

Copyright Notice

Copyright (c) 2015 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. Motivation	3
3. Terminology	4
4. Policy Identifier TLV	4
5. Operation	5
6. Security Considerations	5
7. IANA Considerations	5
8. Acknowledgements	5
9. Normative References	5
Authors' Addresses	6

1. Introduction

[RFC5440] describes the Path Computation Element Protocol (PCEP) for communication between a Path Computation Client (PCC) and a PCE or between a pair of PCEs. [I-D.ietf-pce-stateful-pce] specifies extension to PCEP that allows a PCC to delegate its LSPs to a PCE. The PCE can then update the state of LSPs delegated to it. [I-D.ietf-pce-pce-initiated-lsp] specifies a mechanism allowing a PCE to dynamically instantiate, maintain, and tear down Label Switched Paths (LSPs) without the need for configuring those LSPs on the PCC. Currently, the LSPs can either be signaled via RSVP-TE or can be segment routed as specified in [I-D.ietf-pce-segment-routing].

As described in the next section, a PCEP speaker may want to influence its PCEP counterpart with respect to path selection and other policies. This document describes a PCEP extension to signal policy identifier represented by numeric value using OPTIONAL PCEP

TLV. The specification is applicable to both stateful and stateless PCEP sessions.

2. Motivation

Paths computed using PCEP are subject to various policies on both PCE as well as PCC. For example, in a centralized traffic engineering scenario, network operators may instantiate LSPs and specifies policies for traffic steering, path monitoring, etc., for those LSPs via stateful PCE. Similarly, a PCC can request a path that is diverse from any other path originating from other PCC(s) from a stateful PCE. With a current state of PCEP, introducing such policy requires new PCEP extension. A generic mechanism that allows a PCEP speaker to specify the path policies without the need to know the details of such policies simplifies network operations, avoids frequent software upgrades, as well provides an ability to introduce new policy faster.

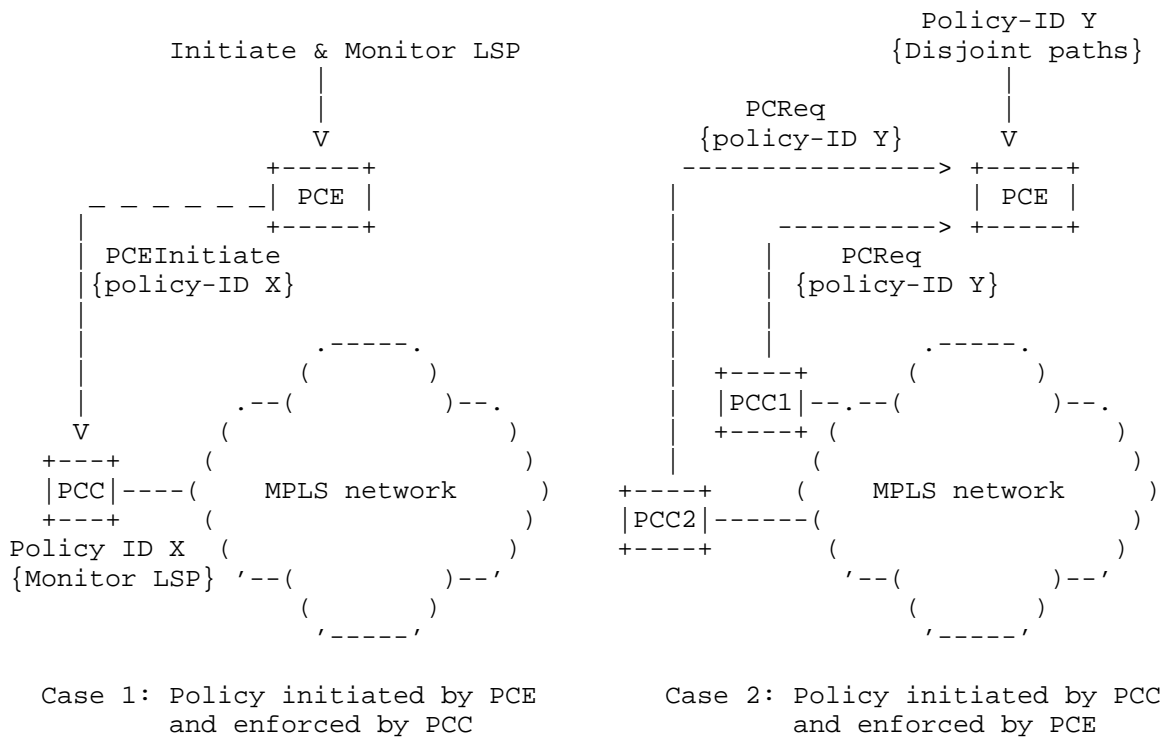


Figure 1: Sample use-cases for carrying policies over PCEP session

3. Terminology

The following terminologies are used in this document:

LSP: Label Switched Path.

PCC: Path Computation Client.

PCE: Path Computation Element

PCEP: Path Computation Element Protocol.

TLV: Type, Length, and Value.

4. Policy Identifier TLV

The new optional TLV is called "POLICY-ID-TLV" whose format is shown in the diagram below is defined to indicate the policies applied to a path. This TLV is associated with the RP or SRP objects specified in[RFC5440] and [I-D.ietf-pce-stateful-pce] respectively. The type of this TLV is to be allocated by IANA.

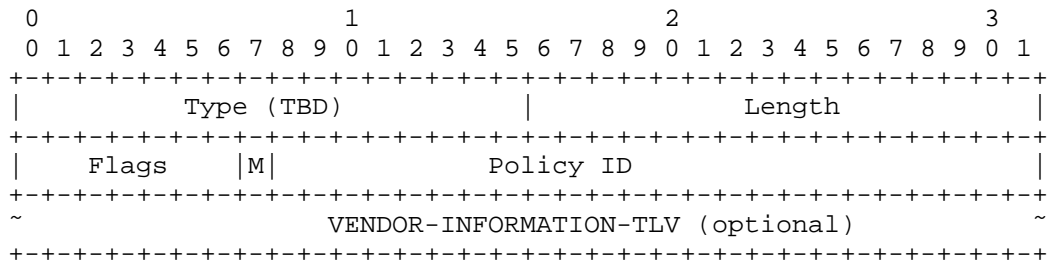


Figure 2: Format of POLICY-ID-TLV

The TLV is formatted according to the rules specified in [RFC5440]. The body of the POLICY-ID-TLV contains one 1-Octet flags and 3-Octet policy identifier. By default, a policy is OPTIONAL. If the M-flag is set, the policy is considered MANDATORY. This TLV can optionally carry vendor-specific information via VENDOR-INFORMATION-TLV whose format and processing rules are specified in [RFC7470]. The presence of VENDOR-INFORMATION-TLV is detected based on the TLV length, and the content and processing rule of vendor-specific information is outside the scope of this specification.

5. Operation

A single message MAY contain more than one POLICY-ID-TLVs. In case, a speaker receives a message containing multiple POLICY-ID-TLVs with the same policy ID, it MUST ignore all except for the first one it encounters in the message. If a PCEP speaker does not recognize the TLV, it MUST ignore the TLV in accordance with ([RFC5440]). If a PCEP speaker recognizes the TLV but does not support a mandatory policy included in the message, it MUST ignore the whole message and send PCErr with Error-Type = 2 (Capability not supported) as well include the POLICY-ID-TLV corresponding to the unsupported policies.

When requesting a path from a PCE using a PCReq message ([RFC5440]), a PCC MAY include the POLICY-ID-TLV in the RP object. The PCE MUST take into account all the policies included in the PCReq otherwise it MUST ignore the whole message and send PCErr message as mentioned above.

In the case of stateful PCE, POLICY-ID-TLV MAY be included in PCReq, PCRpt, PCUpd, and PCInitiate messages as well. When including POLICY-ID-TLV in PCRpt message, the SRP object MUST be present even in cases when the SRP-ID-number is the reserved value of 0x00000000.

6. Security Considerations

No additional security measure is required.

7. IANA Considerations

IANA is requested to allocate a new code point in the PCEP TLV Type Indicators registry, as follows:

Value	Description	Reference
TBD	POLICY-ID-TLV	This document

8. Acknowledgements

9. 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-04 (work in progress), April 2015.

- [I-D.ietf-pce-segment-routing]
Sivabalan, S., Medved, J., Filsfils, C., Crabbe, E.,
Lopez, V., Tantsura, J., Henderickx, W., and J. Hardwick,
"PCEP Extensions for Segment Routing", draft-ietf-pce-
segment-routing-06 (work in progress), August 2015.
- [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-11 (work in progress), April 2015.
- [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>>.
- [RFC4206] Kompella, K. and Y. Rekhter, "Label Switched Paths (LSP)
Hierarchy with Generalized Multi-Protocol Label Switching
(GMPLS) Traffic Engineering (TE)", RFC 4206,
DOI 10.17487/RFC4206, October 2005,
<<http://www.rfc-editor.org/info/rfc4206>>.
- [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>>.
- [RFC5462] Andersson, L. and R. Asati, "Multiprotocol Label Switching
(MPLS) Label Stack Entry: "EXP" Field Renamed to "Traffic
Class" Field", RFC 5462, DOI 10.17487/RFC5462, February
2009, <<http://www.rfc-editor.org/info/rfc5462>>.
- [RFC7470] Zhang, F. and A. Farrel, "Conveying Vendor-Specific
Constraints in the Path Computation Element Communication
Protocol", RFC 7470, DOI 10.17487/RFC7470, March 2015,
<<http://www.rfc-editor.org/info/rfc7470>>.

Authors' Addresses

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

Email: msiva@cisco.com

Clarence Filsfils
Cisco Systems, Inc.
Pegasus Parc
De kleetlaan 6a, DIEGEM BRABANT 1831
BELGIUM

Email: cfilsfil@cisco.com

Jeff Tantsura
Ericsson
300 Holger Way
San Jose, CA 95134
USA

Email: jeff.tantsura@ericsson.com

Jonathan Hardwick
Metaswitch Networks
100 Church Street
Enfield, Middlesex
UK

Email: Jonathan.Hardwick@metaswitch.com

Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: March 7, 2016

N. Wu
Z. Li
Huawei Technologies
September 4, 2015

PCEP Link-State Extensions for Segment Routing
draft-wu-pce-pcep-ls-sr-extension-00

Abstract

This document introduces extensions of PCEP Link-State to export path segment information to a 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 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 March 7, 2016.

Copyright Notice

Copyright (c) 2015 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. PCEP Extensions for Segment Routing	2
2.1. Node Attribute TLVs	2
2.2. Link Attribute TLVs	3
2.3. Prefix Attribute TLVs	3
3. Operational Considerations	4
4. IANA Considerations	4
5. Security Considerations	4
6. References	4
6.1. Normative References	4
6.2. Informative References	5
Authors' Addresses	5

1. Introduction

PCE-LS [I-D.dhodylee-pce-pcep-te-data-extn] is designed to collect topology information and TED across IGP domains. In order to better support SR-TE, this document introduces extensions of PCEP-LS to export path segment information.

2. PCEP Extensions for Segment Routing

PCEP-LS [I-D.dhodylee-pce-pcep-te-data-extn] introduces new message type and new object to accommodate link-state information in PCEP. This document defines new additional TLVs to accommodate segment routing information. The value portion of these new TLVs can reuse the structure defined in [I-D.ietf-isis-segment-routing-extensions] and [I-D.ietf-idr-bgpls-segment-routing-epe].

2.1. Node Attribute TLVs

New

optional, non-transitive node attribute TLVs are defined for carrying segment routing information and are listed as below:

TLV Code Point	Description	Length	Value defined
TBD1	SID/Label Binding	variable	[ISIS-SR]#section2.4
TBD2	SR-Capabilities	variable	[ISIS-SR]#section3.1
TBD3	SR-Algorithm	variable	[ISIS-SR]#section3.2

[ISIS-SR]: <https://datatracker.ietf.org/doc/draft-ietf-isis-segment-routing-extensions/>

Table 1: Node Attribute TLVs

2.2. Link Attribute TLVs

New optional, non-transitive link attribute TLVs are defined for carrying segment routing information and are listed below:

TLV Code Point	Description	Length	Value defined
TBD4	Adjacency Segment	variable	[ISIS-SR]#section2.2.1
TBD5	LAN Adjacency Segment	variable	[ISIS-SR]#section2.2.2
TBD6	Peer Segment	variable	[BGPLS-SR]#section4.3
TBD7	Peer-Set Segment	variable	[BGPLS-SR]#section4.3

[ISIS-SR]: <https://datatracker.ietf.org/doc/draft-ietf-isis-segment-routing-extensions/>

[BGPLS-SR]: <http://datatracker.ietf.org/doc/draft-ietf-idr-bgppls-segment-routing-epe/>

Table 2: Link Attribute TLVs

2.3. Prefix Attribute TLVs

A new optional, non-transitive link attribute TLV is defined for carrying segment routing information and are listed below:

TLV Code Point	Description	Length	Value defined
TBD8	Prefix Segment	variable	[ISIS-SR]#section2.1.2

[ISIS-SR]: <https://datatracker.ietf.org/doc/draft-ietf-isis-segment-routing-extensions/>

Table 3: Prefix Attribute TLVs

3. Operational Considerations

The procedure for segment routing information reporting from PCC to PCE will follow those defined in [I-D.dhodylee-pce-pcep-te-data-extn].

4. IANA Considerations

TBD.

5. Security Considerations

This document does not introduce new security threat.

6. References

6.1. Normative References

- [I-D.dhodylee-pce-pcep-te-data-extn]
Dhody, D., Lee, Y., and D. Ceccarelli, "PCEP Extension for Transporting TE Data", draft-dhodylee-pce-pcep-te-data-extn-02 (work in progress), March 2015.
- [I-D.ietf-idr-bgpls-segment-routing-epe]
Previdi, S., Filsfils, C., Ray, S., Patel, K., Dong, J., and M. Chen, "Segment Routing Egress Peer Engineering BGP-LS Extensions", draft-ietf-idr-bgpls-segment-routing-epe-00 (work in progress), June 2015.
- [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-05 (work in progress), June 2015.
- [I-D.ietf-spring-segment-routing]
Filsfils, C., Previdi, S., Decraene, B., Litkowski, S., and r. rjs@rob.sh, "Segment Routing Architecture", draft-ietf-spring-segment-routing-04 (work in progress), July 2015.
- [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>>.

6.2. Informative References

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

Authors' Addresses

Nan Wu
Huawei Technologies
Huawei Bld., No.156 Beiqing Rd.
Beijing 100095
China

Email: eric.wu@huawei.com

Zhenbin Li
Huawei Technologies
Huawei Bld., No.156 Beiqing Rd.
Beijing 100095
China

Email: lizhenbin@huawei.com