PCE
Internet-Draft
Intended status: Standards Track
Expires: 7 September 2022

R. Chen
ch. Zhu
B. Xu
ZTE Corporation
H. Chen
Futurewei
A. Wang
China Telecom
6 March 2022

PCEP Procedures and Protocol Extensions for Using PCE as a Central Controller (PCECC) of BIER
draft-chen-pce-pcep-extension-pce-controller-bier-03

Abstract

This draft specify a new mechanism where PCE allocates the BIER information centrally and uses PCEP to distribute them to all nodes, then PCC generate a "Bit Index Forwarding Table"(BIFT).

Status of This Memo

Copyright Notice

Table of Contents

1. Introduction

[RFC8283]introduces the architecture for PCE as a central controller as an extension of the architecture described in [RFC4655] and assumes the continued use of PCEP as the protocol used between PCE and PCC. [RFC8283]further examines the motivations and applicability for PCEP as a Southbound Interface (SBI), and introduces the implications for the protocol.

[RFC9050] specify the procedures and PCEP protocol extensions for using the PCE as the central controller for static LSPs, where LSPs can be provisioned as explicit label instructions at each hop on the end-to-end path. Each router along the path must be told what label-forwarding instructions to program and what resources to reserve. The PCE-based controller keeps a view of the network and determines the paths of the end-to-end LSPs, and the controller uses PCEP to communicate with each router along the path of the end-to-end LSP.

[RFC8279] defines a Bit Index Explicit Replication (BIER) architecture where all intended multicast receivers are encoded as a bitmask in the multicast packet header within different encapsulations such as described in [RFC8296]. A router that receives such a packet will forward the packet based on the bit position in the packet header towards the receiver(s) following a precomputed tree for each of the bits in the packet. Each receiver is represented by a unique bit in the bitmask.

In order to reduce the transmission of redundant information, the PCE-based controllers do not allocate the BFIT directly. Instead, the PCC generates the BFIT based on the received bier informations or the node calculates the nexthop by itself. This document specifies the procedures and PCEP protocol extensions when a PCE-based controller is also responsible for configuring the BIER informations.

1.1. Requirements Language

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

2. PCECC BIER Requirements

Following key requirements for PCECC-BIER should be considered when‘designing the PCECC based solution:

* PCEP speaker supporting this draft needs to have the capability to advertise its PCECC-BIER capability to its peers.

* PCEP speaker not supporting this draft needs to be able to reject PCECC-BIER related message with a reason code that indicates no support for PCECC.

* PCEP procedures needs to provide a means to update (or cleanup) the BIER related informations (BIER subdomain-id, BFR-id and BSL etc) to the PCC.

* PCEP procedures needs to provide a means to update (or cleanup) the "Bit Index Forwarding Table"(BIFT) to the PCC.

* PCEP procedures needs to provide a means to synchronize the BIER related informations (BIER subdomain-id, BFR-id and BSL etc) between PCE to PCC in the PCEP messages.

3. Procedures for Using the PCE as the Central Controller (PCECC) in BIER

Active stateful PCE is described in [RFC8231]. PCE as a central controller (PCECC) reuses existing Active stateful PCE mechanism as much as possible to control the LSP.

This document uses the same PCEP messages and its extensions which are described in [RFC9050] for PCECC-BIER as well.

PCEP messages PCRpt, PCInitiate, PCUpd are also used to send LSP Reports, LSP setup and LSP update respectively. The extended PCInitiate message described in [RFC9050] is used to download or cleanup central controller’s instructions (CCIs) (BIER related informations and "Bit Index Forwarding Table"(BIFT) in scope of this document). The extended PCRpt message described in [RFC9050] is also used to report the CCIs (BIER related informations) from PCC to PCE.

[RFC9050] specify an object called CCI for the encoding of central controller’s instructions. This document extends the CCI by defining another object-type for BIER.

3.1. PCECC Capability Advertisement

During PCEP Initialization Phase, PCEP Speakers (PCE or PCC) advertise their support of PCECC extensions. A PCEP Speaker includes the "PCECC Capability" sub-TLV, described in [RFC9050].

This document adds B-bit in PCECC-CAPABILITY sub-TLV for BIER.

3.2. New BIER Path Setup

The PCEP messages pertaining to PCECC-BIER MUST include PATH-SETUP-TYPE TLV [RFC8408] with PST=TBD in the SRP object to clearly identify the PCECC-BIER is intended.

3.3. PCECC BIER information allocation and Generation of BFIT

There are two ways to generate a "Bit Index Forwarding Table"(BIFT):

* The PCECC allocate parameters(BIER subdomain-id, BFR-id, BAR and IPA) carried by CCI object, parameters(BFR prefix, BSL, Encapsulation Type, BIFT ID, and Max SI) carried by BIER Encapsulation TLV and parameters (BFR prefix)carried by OFEC Object to the PCC. On receiving the BIER informations allocation, each node (PCC) uses IGP protocol to distribute BIER related information to other nodes. The node calculate the nexthop. In this case, Each node (PCC) only needs to be allocated its own BIER informations by the PCECC.

* In scenarios where the IGP protocol is not used/available,Each node (PCC) is allocated its own and neighbor BIER informations by the PCECC, then PCC generates a BIFT based on the informations it receives. The BIER informations include BIER subdomain-id and BFR-id carried by CCI object, BFR prefix, BSL, Encapsulation Type, BIFT ID, and Max SI carried by BIER Encapsulation TLV , BFR-NBR carried by Address TLV and BFR prefix carried by OFEC Object. The BIFT mainly includes BFR ID, F-BM and BFR nexthop.

3.4. Redundant PCEs

[I-D.ietf-pce-state-sync] describes synchronization mechanism between the stateful PCEs. The BIER informations allocated by a PCE MUST also be synchronized among PCEs for PCECC BIER state synchronization.

3.5. Re Delegation and Cleanup

[RFC9050] describes the action needed for CCIs for the Basic PCECC LSP on this terminated session. Similarly actions should be applied for the BIER information as well.

3.6. Synchronization of BIER information Allocations

[RFC9050] describes the synchronization of Central Controller's Instructions (CCI) via LSP state synchronization as described in [RFC8231] and [RFC8232].Same procedures should be applied for BIER informations as well.

4. PCEP extension

4.1. The OPEN Object

4.1.1. PCECC Capability sub-TLV

[RFC9050] defined the PCECC-CAPABILITY TLV. A new B-bit is defined in PCECC-CAPABILITY sub-TLV for PCECC-BIER:

```
 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=TBD                  |               Length          |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                Flags                                    |B|I|S|
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 1

where:

B (PCECC-BIER-CAPABILITY - 1 bit): If set to 1 by a PCEP speaker, it indicates that the PCEP speaker is capable for PCECC-BIER capability and PCE would allocate BIER information on this session.

4.2. PATH-SETUP-TYPE TLV

The PATH-SETUP-TYPE TLV is defined in [RFC8408]. PST = TBD is used when Path is setup via PCECC BIER mode.On a PCRpt/PCUpd/PCInitiate message, the PST=TBD indicates that this path was setup via a PCECC-BIER based mechanism where either the BIER informations and BIER forwarding entries were allocated/instructed by PCE via PCECC mechanism.

4.3. CCI object

The Central Control Instructions (CCI) Object is used by the PCE to specify the forwarding instructions is defined in [RFC9050]. This document defines another object-type for BIER purpose.

CCI Object-Type is TBD for BIER as below

```
 0                   1                   2                   3
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                             CC-ID                             |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
| subdomain-ID  |     BAR       |     IPA       | Flags         |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|           BFR-ID              |        Reserved               |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                                                               |
//                       Optional TLV                          //
|                                                               |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 2

where:

The field CC-ID is as described in [RFC9050].

BIER subdomain-ID: Unique value identifying the BIER subdomain. (as defined in [RFC8401].

BAR:BIER Algorithm, as documented in [RFC8401]. Specifies a BIER-specific algorithm used to calculate underlay paths to reach BFERs. Values are allocated from the "BIER Algorithms" registry.

IPA:IGP Algorithm, as documented in [RFC8401]. Specifies an IGP Algorithm to either modify, enhance, or replace the calculation of underlay paths to reach BFERs as defined by the BAR value. Values are from the IGP Algorithm registry. 1 octet.

Flags (16 bit): A field used to carry any additional information pertaining to the CCI.

BFR-ID: A 2-octet field encoding the BFR-id, as documented in [RFC8279].

Optional TLV: There are two optional TLV are defined/reused in this draft.

4.3.1. BIER Encapsulation Sub TLV

```
 0                   1                   2                   3
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|             Type              |             Length            |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|         Flags             | ET|             Reserved          |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|   Max SI      |BS Len |                  BIFT-id              |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 3

where:

The code point for the TLV type is to be defined by IANA.

Length:4

ET-Flag:ET(Encapsulation type) Flag,There are two Encapsulation Types:

* 0b00-MPLS encapsulation.

* 0b01-Non-MPLS encapsulation.

Max SI: A 1 octet field encoding the Maximum Set Identifier(Section 1 of [RFC8279]) used in the encapsulation for this BIER subdomain for this BitString length.

Local BitString Length (BS Len): Encoded BitString length as per [RFC8296].

BIFT-id: A 20 bit field encoding the first BIFT-id of the BIFT-id range.

4.3.2. Address TLVs

Address TLVs described in [RFC9050] are used to associate the next-hop information, so we Reuse ADDRESS TLV to carry the BFR out-interface and nexthop informations.

4.4. FEC Object

BIER information is always associated with a host prefix, so we reuse FEC Object 1'IPv4 Node ID' and FEC Object-Type 2 'IPv6 Node ID' defined in [RFC8664] to carry the BFR prefix.

5. Acknowledgements

We would like to thank Dhruv Dhody for their useful comments and suggestions.

6. IANA Considerations

TBD.

7. Security Considerations

The PCECC extension are based on the existing PCEP messages and thus the security considerations described in

The PCECC extension are based on the existing PCEP messages and thus the security considerations described in [RFC5440] ,[RFC8231] ,[RFC8281], and [RFC9050] apply to this draft.

8. Normative References

[I-D.ietf-pce-state-sync]
Litkowski, S., Sivabalan, S., Li, C., and H. Zheng, "Inter Stateful Path Computation Element (PCE) Communication Procedures.", Work in Progress, Internet-Draft, draft-ietf-pce-state-sync-01, 20 October 2021, <https://datatracker.ietf.org/doc/html/draft-ietf-pce-state-sync-01>.

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <https://www.rfc-editor.org/info/rfc2119>.

[RFC4655] Farrel, A., Vasseur, J.-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>.

[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>.

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

[RFC8232] Crabbe, E., Minei, I., Medved, J., Varga, R., Zhang, X., and D. Dhody, "Optimizations of Label Switched Path State Synchronization Procedures for a Stateful PCE", RFC 8232, DOI 10.17487/RFC8232, September 2017, <https://www.rfc-editor.org/info/rfc8232>.

[RFC8279] Wijnands, IJ., Ed., Rosen, E., Ed., Dolganow, A., Przygienda, T., and S. Aldrin, "Multicast Using Bit Index Explicit Replication (BIER)", RFC 8279, DOI 10.17487/RFC8279, November 2017, <https://www.rfc-editor.org/info/rfc8279>.

[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>.

[RFC8296] Wijnands, IJ., Ed., Rosen, E., Ed., Dolganow, A., Tantsura, J., Aldrin, S., and I. Meilik, "Encapsulation for Bit Index Explicit Replication (BIER) in MPLS and Non-MPLS Networks", RFC 8296, DOI 10.17487/RFC8296, January 2018, <https://www.rfc-editor.org/info/rfc8296>.

[RFC8401] Ginsberg, L., Ed., Przygienda, T., Aldrin, S., and Z. Zhang, "Bit Index Explicit Replication (BIER) Support via IS-IS", RFC 8401, DOI 10.17487/RFC8401, June 2018, <https://www.rfc-editor.org/info/rfc8401>.

[RFC8408] Sivabalan, S., Tantsura, J., Minei, I., Varga, R., and J. Hardwick, "Conveying Path Setup Type in PCE Communication Protocol (PCEP) Messages", RFC 8408, DOI 10.17487/RFC8408, July 2018, <https://www.rfc-editor.org/info/rfc8408>.

[RFC8664] Sivabalan, S., Filsfils, C., Tantsura, J., Henderickx, W., and J. Hardwick, "Path Computation Element Communication Protocol (PCEP) Extensions for Segment Routing", RFC 8664, DOI 10.17487/RFC8664, December 2019, <https://www.rfc-editor.org/info/rfc8664>.

[RFC9050] Li, Z., Peng, S., Negi, M., Zhao, Q., and C. Zhou, "Path Computation Element Communication Protocol (PCEP) Procedures and Extensions for Using the PCE as a Central Controller (PCECC) of LSPs", RFC 9050, DOI 10.17487/RFC9050, July 2021, <https://www.rfc-editor.org/info/rfc9050>.

Authors' Addresses

Ran Chen
ZTE Corporation
Nanjing
China
Email: chen.ran@zte.com.cn

Chun Zhu
ZTE Corporation
Nanjing
China

Email: zhu.chun@zte.com.cn

BenChong Xu
ZTE Corporation
Nanjing
China
Email: xu.benchong@zte.com.cn

Huaimo Chen
Futurewei
Nanjing,
United States of America
Email: Huaimo.chen@futurewei.com

Aijun Wang
China Telecom
Nanjing
China
Email: wangaj3@chinatelecom.cn

PCE
Internet-Draft
Intended status: Standards Track
Expires: 22 April 2024

R. Chen
ch. Zhu
B. Xu
ZTE Corporation
H. Chen
Futurewei
A. Wang
China Telecom
20 October 2023

PCEP Procedures and Protocol Extensions for Using PCE as a Central Controller (PCECC) of BIER
draft-chen-pce-pcep-extension-pce-controller-bier-05

Abstract

This draft specify a new mechanism where PCE allocates the BIER information centrally and uses PCEP to distribute them to all nodes, then PCC generate a "Bit Index Forwarding Table"(BIFT).

Status of This Memo

Copyright Notice

Table of Contents

1. Introduction

[RFC8283]introduces the architecture for PCE as a central controller as an extension of the architecture described in [RFC4655] and assumes the continued use of PCEP as the protocol used between PCE and PCC. [RFC8283]further examines the motivations and applicability for PCEP as a Southbound Interface (SBI), and introduces the implications for the protocol.

[RFC9050] specify the procedures and PCEP protocol extensions for using the PCE as the central controller for static LSPs, where LSPs can be provisioned as explicit label instructions at each hop on the end-to-end path. Each router along the path must be told what label-forwarding instructions to program and what resources to reserve. The PCE-based controller keeps a view of the network and determines the paths of the end-to-end LSPs, and the controller uses PCEP to communicate with each router along the path of the end-to-end LSP.

[RFC8279] defines a Bit Index Explicit Replication (BIER) architecture where all intended multicast receivers are encoded as a bitmask in the multicast packet header within different encapsulations such as described in [RFC8296]. A router that receives such a packet will forward the packet based on the bit position in the packet header towards the receiver(s) following a precomputed tree for each of the bits in the packet. Each receiver is represented by a unique bit in the bitmask.

In order to reduce the transmission of redundant information, the PCE-based controllers do not allocate the BFIT directly. Instead, the PCC generates the BFIT based on the received bier informations or the node calculates the nexthop by itself. This document specifies the procedures and PCEP protocol extensions when a PCE-based controller is responsible for configuring the BIER informations.

1.1. Requirements Language

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

2. PCECC BIER Requirements

Following key requirements for PCECC-BIER should be considered when‘designing the PCECC based solution:

* PCEP speaker supporting this draft needs to have the capability to advertise its PCECC-BIER capability to its peers.

* PCEP speaker not supporting this draft needs to be able to reject PCECC-BIER related message with a reason code that indicates no support for PCECC.

* PCEP procedures needs to provide a means to update (or cleanup) the BIER related informations (BIER subdomain-id, BFR-id and BSL etc) to the PCC.

* PCEP procedures needs to provide a means to update (or cleanup) the "Bit Index Forwarding Table"(BIFT) to the PCC.

* PCEP procedures needs to provide a means to synchronize the BIER related informations (BIER subdomain-id, BFR-id and BSL etc) between PCE to PCC in the PCEP messages.

3. Procedures for Using the PCE as the Central Controller (PCECC) in BIER

Active stateful PCE is described in [RFC8231]. PCE as a central controller (PCECC) reuses existing Active stateful PCE mechanism as much as possible to control the LSP.

This document uses the same PCEP messages and its extensions which are described in [RFC9050] for PCECC-BIER as well.

PCEP messages PCRpt, PCInitiate, PCUpd are also used to send LSP Reports, LSP setup and LSP update respectively. The extended PCInitiate message described in [RFC9050] is used to download or cleanup central controller’s instructions (CCIs) (BIER related informations and "Bit Index Forwarding Table"(BIFT) in scope of this document). The extended PCRpt message described in [RFC9050] is also used to report the CCIs (BIER related informations) from PCC to PCE.

[RFC9050] specify an object called CCI for the encoding of central controller’s instructions. This document extends the CCI by defining another object-type for BIER.

3.1. PCECC Capability Advertisement

During PCEP Initialization Phase, PCEP Speakers (PCE or PCC) advertise their support of PCECC extensions. A PCEP Speaker includes the "PCECC Capability" sub-TLV, described in [RFC9050].

This document adds B-bit in PCECC-CAPABILITY sub-TLV for BIER.

3.2. New BIER Path Setup

The PCEP messages pertaining to PCECC-BIER MUST include PATH-SETUP-TYPE TLV [RFC8408] with PST=TBD in the SRP object to clearly identify the PCECC-BIER is intended.

3.3. PCECC BIER information allocation and Generation of BFIT

There are two ways to generate a "Bit Index Forwarding Table"(BIFT):

* The PCECC allocate parameters(BIER subdomain-id, BFR-id, BAR and IPA) carried by CCI object, parameters(BFR prefix, BSL, Encapsulation Type, BIFT ID, and Max SI) carried by BIER Encapsulation TLV and parameters (BFR prefix)carried by OFEC Object to the PCC. On receiving the BIER informations allocation, each node (PCC) uses IGP protocol to distribute BIER related information to other nodes. The node calculate the nexthop. In this case, Each node (PCC) only needs to be allocated its own BIER informations by the PCECC.

* In scenarios where the IGP protocol is not used/available,Each node (PCC) is allocated its own and neighbor BIER informations by the PCECC, then PCC generates a BIFT based on the informations it receives. The BIER informations include BIER subdomain-id and BFR-id carried by CCI object, BFR prefix, BSL, Encapsulation Type, BIFT ID, and Max SI carried by BIER Encapsulation TLV , BFR-NBR carried by Address TLV and BFR prefix carried by OFEC Object. The BIFT mainly includes BFR ID, F-BM and BFR nexthop.

3.4. Redundant PCEs

[I-D.ietf-pce-state-sync] describes synchronization mechanism between the stateful PCEs. The BIER informations allocated by a PCE MUST also be synchronized among PCEs for PCECC BIER state synchronization.

3.5. Re Delegation and Cleanup

[RFC9050] describes the action needed for CCIs for the Basic PCECC LSP on this terminated session. Similarly actions should be applied for the BIER information as well.

3.6. Synchronization of BIER information Allocations

[RFC9050] describes the synchronization of Central Controller's Instructions (CCI) via LSP state synchronization as described in [RFC8231] and [RFC8232].Same procedures should be applied for BIER informations as well.

4. PCEP extension

4.1. The OPEN Object

4.1.1. PCECC Capability sub-TLV

[RFC9050] defined the PCECC-CAPABILITY TLV. A new B-bit is defined in PCECC-CAPABILITY sub-TLV for PCECC-BIER:

```
 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=TBD                  |               Length          |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                Flags                                        |B|I|S|
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 1

where:

B (PCECC-BIER-CAPABILITY - 1 bit): If set to 1 by a PCEP speaker, it indicates that the PCEP speaker is capable for PCECC-BIER capability and PCE would allocate BIER information on this session.

4.2. PATH-SETUP-TYPE TLV

The PATH-SETUP-TYPE TLV is defined in [RFC8408]. PST = TBD is used when Path is setup via PCECC BIER mode.On a PCRpt/PCUpd/PCInitiate message, the PST=TBD indicates that this path was setup via a PCECC-BIER based mechanism where either the BIER informations and BIER forwarding entries were allocated/instructed by PCE via PCECC mechanism.

4.3. CCI object

The Central Control Instructions (CCI) Object is used by the PCE to specify the forwarding instructions is defined in [RFC9050]. This document defines another object-type for BIER purpose.

CCI Object-Type is TBD for BIER as below

```
 0                   1                   2                   3
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                             CC-ID                             |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
| subdomain-ID  |     BAR       |     IPA       | Flags         |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|           BFR-ID              |        Reserved               |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                                                               |
//                       Optional TLV                          //
|                                                               |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 2

where:

The field CC-ID is as described in [RFC9050].

BIER subdomain-ID: Unique value identifying the BIER subdomain. (as defined in [RFC8401].

BAR:BIER Algorithm, as documented in [RFC8401]. Specifies a BIER-specific algorithm used to calculate underlay paths to reach BFERs. Values are allocated from the "BIER Algorithms" registry.

IPA:IGP Algorithm, as documented in [RFC8401]. Specifies an IGP Algorithm to either modify, enhance, or replace the calculation of underlay paths to reach BFERs as defined by the BAR value. Values are from the IGP Algorithm registry. 1 octet.

Flags (16 bit): A field used to carry any additional information pertaining to the CCI.

BFR-ID: A 2-octet field encoding the BFR-id, as documented in [RFC8279].

Optional TLV: There are two optional TLV are defined/reused in this draft.

4.3.1. BIER Encapsulation Sub TLV

```
 0                   1                   2                   3
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|             Type              |             Length            |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|         Flags             | ET|             Reserved          |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|   Max SI      |BS Len |                  BIFT-id              |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 3

where:

The code point for the TLV type is to be defined by IANA.

Length:4

ET-Flag:ET(Encapsulation type) Flag,There are two Encapsulation Types:

* 0b00-MPLS encapsulation.

* 0b01-Non-MPLS encapsulation.

Max SI: A 1 octet field encoding the Maximum Set Identifier(Section 1 of [RFC8279]) used in the encapsulation for this BIER subdomain for this BitString length.

Local BitString Length (BS Len): Encoded BitString length as per [RFC8296].

BIFT-id: A 20 bit field encoding the first BIFT-id of the BIFT-id range.

4.3.2. Address TLVs

Address TLVs described in [RFC9050] are used to associate the next-hop information, so we Reuse ADDRESS TLV to carry the BFR out-interface and nexthop informations.

4.4. FEC Object

BIER information is always associated with a host prefix, so we reuse FEC Object 1'IPv4 Node ID' and FEC Object-Type 2 'IPv6 Node ID' defined in [RFC8664] to carry the BFR prefix.

5. Acknowledgements

We would like to thank Dhruv Dhody for their useful comments and suggestions.

6. IANA Considerations

TBD.

7. Security Considerations

The PCECC extension are based on the existing PCEP messages and thus the security considerations described in

The PCECC extension are based on the existing PCEP messages and thus the security considerations described in [RFC5440] ,[RFC8231] ,[RFC8281], and [RFC9050] apply to this draft.

8. Normative References

[I-D.ietf-pce-state-sync]
Litkowski, S., Sivabalan, S., Li, C., and H. Zheng, "Inter Stateful Path Computation Element (PCE) Communication Procedures.", Work in Progress, Internet-Draft, draft-ietf-pce-state-sync-05, 9 July 2023, <https://datatracker.ietf.org/doc/html/draft-ietf-pce-state-sync-05>.

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <https://www.rfc-editor.org/info/rfc2119>.

[RFC4655] Farrel, A., Vasseur, J.-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>.

[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>.

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

[RFC8232] Crabbe, E., Minei, I., Medved, J., Varga, R., Zhang, X., and D. Dhody, "Optimizations of Label Switched Path State Synchronization Procedures for a Stateful PCE", RFC 8232, DOI 10.17487/RFC8232, September 2017, <https://www.rfc-editor.org/info/rfc8232>.

[RFC8279] Wijnands, IJ., Ed., Rosen, E., Ed., Dolganow, A., Przygienda, T., and S. Aldrin, "Multicast Using Bit Index Explicit Replication (BIER)", RFC 8279, DOI 10.17487/RFC8279, November 2017, <https://www.rfc-editor.org/info/rfc8279>.

[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>.

[RFC8296] Wijnands, IJ., Ed., Rosen, E., Ed., Dolganow, A., Tantsura, J., Aldrin, S., and I. Meilik, "Encapsulation for Bit Index Explicit Replication (BIER) in MPLS and Non-MPLS Networks", RFC 8296, DOI 10.17487/RFC8296, January 2018, <https://www.rfc-editor.org/info/rfc8296>.

[RFC8401] Ginsberg, L., Ed., Przygienda, T., Aldrin, S., and Z. Zhang, "Bit Index Explicit Replication (BIER) Support via IS-IS", RFC 8401, DOI 10.17487/RFC8401, June 2018, <https://www.rfc-editor.org/info/rfc8401>.

[RFC8408] Sivabalan, S., Tantsura, J., Minei, I., Varga, R., and J. Hardwick, "Conveying Path Setup Type in PCE Communication Protocol (PCEP) Messages", RFC 8408, DOI 10.17487/RFC8408, July 2018, <https://www.rfc-editor.org/info/rfc8408>.

[RFC8664] Sivabalan, S., Filsfils, C., Tantsura, J., Henderickx, W., and J. Hardwick, "Path Computation Element Communication Protocol (PCEP) Extensions for Segment Routing", RFC 8664, DOI 10.17487/RFC8664, December 2019, <https://www.rfc-editor.org/info/rfc8664>.

[RFC9050] Li, Z., Peng, S., Negi, M., Zhao, Q., and C. Zhou, "Path Computation Element Communication Protocol (PCEP) Procedures and Extensions for Using the PCE as a Central Controller (PCECC) of LSPs", RFC 9050, DOI 10.17487/RFC9050, July 2021, <https://www.rfc-editor.org/info/rfc9050>.

Authors' Addresses

Ran Chen
ZTE Corporation
Nanjing
China
Email: chen.ran@zte.com.cn

Chun Zhu
ZTE Corporation
Nanjing
China

Email: zhu.chun@zte.com.cn

BenChong Xu
ZTE Corporation
Nanjing
China
Email: xu.benchong@zte.com.cn

Huaimo Chen
Futurewei
Nanjing,
United States of America
Email: Huaimo.chen@futurewei.com

Aijun Wang
China Telecom
Nanjing
China
Email: wangaj3@chinatelecom.cn

PCE
Internet-Draft
Intended status: Standards Track
Expires: August 8, 2022

H. Yuan
UnionPay
T. Zhou
W. Li
G. Fioccola
Y. Wang
Huawei

February 4, 2022

Path Computation Element Communication Protocol (PCEP) Extensions to Enable IFIT
draft-chen-pce-pcep-ifit-06

Abstract

This document defines PCEP extensions to distribute In-situ Flow Information Telemetry (IFIT) information. So that IFIT behavior can be enabled automatically when the path is instantiated. In-situ Flow Information Telemetry (IFIT) refers to network OAM data plane on-path telemetry techniques, in particular the most popular are In-situ OAM (IOAM) and Alternate Marking. The IFIT attributes here described can be generalized for all path types but the application to Segment Routing (SR) is considered in this document. This document extends PCEP to carry the IFIT attributes under the stateful PCE model.

Requirements Language

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

Status of This Memo

This Internet-Draft will expire on August 8, 2022.

Copyright Notice

Table of Contents

1. Introduction

In-situ Flow Information Telemetry (IFIT) refers to network OAM (Operations, Administration, and Maintenance) data plane on-path telemetry techniques, including In-situ OAM (IOAM) [I-D.ietf-ippm-ioam-data] and Alternate Marking [RFC8321]. It can provide flow information on the entire forwarding path on a per-packet basis in real time.

An automatic network requires the Service Level Agreement (SLA) monitoring on the deployed service. So that the system can quickly detect the SLA violation or the performance degradation, hence to change the service deployment.

This document defines extensions to PCEP to distribute paths carrying IFIT information. So that IFIT behavior can be enabled automatically when the path is instantiated.

RFC 5440 [RFC5440] describes the Path Computation Element Protocol (PCEP) as a communication mechanism between a Path Computation Client (PCC) and a Path Computation Element (PCE), or between a PCE and a PCE.

RFC 8231 [RFC8231] specifies extensions to PCEP to enable stateful control and it describes two modes of operation: passive stateful PCE and active stateful PCE. Further, RFC 8281 [RFC8281] describes the setup, maintenance, and teardown of PCE-initiated LSPs for the stateful PCE model.

When a PCE is used to initiate paths using PCEP, it is important that the head end of the path also understands the IFIT behavior that is intended for the path. When PCEP is in use for path initiation it makes sense for that same protocol to be used to also carry the IFIT attributes that describe the IOAM or Alternate Marking procedure that needs to be applied to the data that flow those paths.

The PCEP extension defined in this document allows to signal the IFIT capabilities. In this way IFIT methods are automatically activated and running. The flexibility and dynamicity of the IFIT applications are given by the use of additional functions on the controller and on the network nodes, but this is out of scope here.

IFIT is a solution focusing on network domains according to [RFC8799] that introduces the concept of specific domain solutions. A network domain consists of a set of network devices or entities within a single administration. As mentioned in [RFC8799], for a number of reasons, such as policies, options supported, style of network management and security requirements, it is suggested to limit

applications including the emerging IFIT techniques to a controlled domain. Hence, the IFIT methods MUST be typically deployed in such controlled domains.

The Use Case of Segment Routing (SR) is also discussed considering that IFIT methods are becoming mature for Segment Routing over the MPLS data plane (SR-MPLS) and Segment Routing over IPv6 data plane (SRv6). SR policy [I-D.ietf-spring-segment-routing-policy] is a set of candidate SR paths consisting of one or more segment lists and necessary path attributes. It enables instantiation of an ordered list of segments with a specific intent for traffic steering. The PCEP extension defined in this document also enables SR policy with native IFIT, that can facilitate the closed loop control and enable the automation of SR service.

It is to be noted the companion document [I-D.qin-idr-sr-policy-ifit] that proposes the BGP extension to enable IFIT methods for SR policy.

2. PCEP Extensions for IFIT Attributes

This document is to add IFIT attribute TLVs as PCEP Extensions. The following sections will describe the requirement and usage of different IFIT modes, and define the corresponding TLV encoding in PCEP.

The IFIT attributes here described can be generalized and included as TLVs carried inside the LSPA (LSP Attributes) object in order to be applied for all path types, as long as they support the relevant data plane telemetry method. IFIT Attributes TLVs are optional and can be taken into account by the PCE during path computation and by the PCC during path setup. In general, the LSPA object can be carried within a PCInitiate message, a PCUpd message, or a PCRpt message in the stateful PCE model.

In this document it is considered the case of SR Policy since IOAM and Alternate Marking are more mature especially for Segment Routing (SR) and for IPv6.

It is to be noted that, if it is needed to apply different IFIT methods for each Segment List, the IFIT attributes can be added into the PATH-ATTRIB object, instead of the LSPA object, according to [I-D.koldychev-pce-multipath] that defines PCEP Extensions for Signaling Multipath Information.

2.1. IFIT for SR Policies

RFC 8664 [RFC8664] and [I-D.ietf-pce-segment-routing-ipv6] specify extensions to the Path Computation Element Communication Protocol (PCEP) that allow a stateful PCE to compute and initiate Traffic-Engineering (TE) paths, as well as a Path Computation Client (PCC) to request a path subject to certain constraints and optimization criteria in SR networks both for SR-MPLS and SRv6.

IFIT attibutes, here defined as TLVs for the LSPA object, complement both RFC 8664 [RFC8664], [I-D.ietf-pce-segment-routing-ipv6] and [I-D.ietf-pce-segment-routing-policy-cp].

3. IFIT capability advertisement TLV

During the PCEP initialization phase, PCEP speakers (PCE or PCC) SHOULD advertise their support of IFIT methods (e.g. IOAM and Alternate Marking).

A PCEP speaker includes the IFIT-CAPABILITY TLVs in the OPEN object to advertise its support for PCEP IFIT extensions. The presence of the IFIT-CAPABILITY TLV in the OPEN object indicates that the IFIT methods are supported.

RFC 8664 [RFC8664] and [I-D.ietf-pce-segment-routing-ipv6] define a new Path Setup Type (PST) for SR and also define the SR-PCE-CAPABILITY sub-TLV. This document defined a new IFIT-CAPABILITY TLV, that is an optional TLV for use in the OPEN Object for IFIT attributes via PCEP capability advertisement.

```
 0                   1                   2                   3
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|           Type                |            Length=4           |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                            Flags                    |P|I|D|E|M|
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Fig. 1 IFIT-CAPABILITY TLV Format

Where:

Type: to be assigned by IANA.

Length: 4.

Flags: The following flags are defined in this document:

P: IOAM Pre-allocated Trace Option Type-enabled flag [I-D.ietf-ippm-ioam-data]. If set to 1 by a PCC, the P flag indicates that the PCC allows instantiation of the IOAM Pre-allocated Trace feature by a PCE. If set to 1 by a PCE, the P flag indicates that the PCE supports the IOAM Pre-allocated Trace feature instantiation. The P flag MUST be set by both PCC and PCE in order to support the IOAM Pre-allocated Trace instantiation

I: IOAM Incremental Trace Option Type-enabled flag [I-D.ietf-ippm-ioam-data]. If set to 1 by a PCC, the I flag indicates that the PCC allows instantiation of the IOAM Incremental Trace feature by a PCE. If set to 1 by a PCE, the I flag indicates that the PCE supports the relative IOAM Incremental Trace feature instantiation. The I flag MUST be set by both PCC and PCE in order to support the IOAM Incremental Trace feature instantiation

D: IOAM DEX Option Type-enabled flag [I-D.ietf-ippm-ioam-direct-export]. If set to 1 by a PCC, the D flag indicates that the PCC allows instantiation of the relative IOAM DEX feature by a PCE. If set to 1 by a PCE, the D flag indicates that the PCE supports the relative IOAM DEX feature instantiation. The D flag MUST be set by both PCC and PCE in order to support the IOAM DEX feature instantiation

E: IOAM E2E Option Type-enabled flag [I-D.ietf-ippm-ioam-data]. If set to 1 by a PCC, the E flag indicates that the PCC allows instantiation of the relative IOAM E2E feature by a PCE. If set to 1 by a PCE, the E flag indicates that the PCE supports the relative IOAM E2E feature instantiation. The E flag MUST be set by both PCC and PCE in order to support the IOAM E2E feature instantiation

M: Alternate Marking enabled flag RFC 8321 [RFC8321]. If set to 1 by a PCC, the M flag indicates that the PCC allows instantiation of the relative Alternate Marking feature by a PCE. If set to 1 by a PCE, the M flag indicates that the PCE supports the relative Alternate Marking feature instantiation. The M flag MUST be set by both PCC and PCE in order to support the Alternate Marking feature instantiation

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

Advertisement of the IFIT-CAPABILITY TLV implies support of IFIT methods (IOAM and/or Alternate Marking) as well as the objects, TLVs, and procedures defined in this document. It is worth mentioning that IOAM and Alternate Marking can be activated one at a time or can

coexist; so it is possible to have only IOAM or only Alternate Marking enabled but they are recognized in general as IFIT capability.

The IFIT Capability Advertisement can imply the following cases:

- o The PCEP protocol extensions for IFIT MUST NOT be used if one or both PCEP speakers have not included the IFIT-CAPABILITY TLV in their respective OPEN message.

- o A PCEP speaker that does not recognize the extensions defined in this document would simply ignore the TLVs as per RFC 5440 [RFC5440].

- o If a PCEP speaker supports the extensions defined in this document but did not advertise this capability, then upon receipt of IFIT-ATTRIBUTES TLV in the LSP Attributes (LSPA) object, it SHOULD generate a PCErr with Error-Type 19 (Invalid Operation) with the relative Error-value "IFIT capability not advertised" and ignore the IFIT-ATTRIBUTES TLV.

4. IFIT Attributes TLV

The IFIT-ATTRIBUTES TLV provides the configurable knobs of the IFIT feature, and it can be included as an optional TLV in the LSPA object (as described in RFC 5440 [RFC5440]).

For a PCE-initiated LSP RFC 8281 [RFC8281], this TLV is included in the LSPA object with the PCInitiate message. For the PCC-initiated delegated LSPs, this TLV is carried in the Path Computation State Report (PCRpt) message in the LSPA object. This TLV is also carried in the LSPA object with the Path Computation Update Request (PCUpd) message to direct the PCC (LSP head-end) to make updates to IFIT attributes.

The TLV is encoded in all PCEP messages for the LSP if IFIT feature is enabled. The absence of the TLV indicates the PCEP speaker wishes to disable the feature. This TLV includes multiple IFIT-ATTRIBUTES sub-TLVs. The IFIT-ATTRIBUTES sub-TLVs are included if there is a change since the last information sent in the PCEP message. The default values for missing sub-TLVs apply for the first PCEP message for the LSP.

The format of the IFIT-ATTRIBUTES TLV is shown in the following figure:

```
 0                   1                   2                   3
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|             Type              |           Length              |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                                                               |
//                           sub-TLVs                          //
|                                                               |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Fig. 2 IFIT-ATTRIBUTES TLV Format

Where:

Type: to be assigned by IANA.

Length: The Length field defines the length of the value portion in bytes as per RFC 5440 [RFC5440].

Value: This comprises one or more sub-TLVs.

The following sub-TLVs are defined in this document:

Type	Len	Name
1	8	IOAM Pre-allocated Trace Option
2	8	IOAM Incremental Trace Option
3	12	IOAM Directly Export Option
4	4	IOAM Edge-to-Edge Option
5	4	Enhanced Alternate Marking

Fig. 3 Sub-TLV Types of the IFIT-ATTRIBUTES TLV

4.1. IOAM Sub-TLVs

In-situ Operations, Administration, and Maintenance (IOAM) [I-D.ietf-ippm-ioam-data] records operational and telemetry information in the packet while the packet traverses a path between two points in the network. In terms of the classification given in RFC 7799 [RFC7799] IOAM could be categorized as Hybrid Type 1. IOAM mechanisms can be leveraged where active OAM do not apply or do not offer the desired results.

For the SR use case, when SR policy enables IOAM, the IOAM header will be inserted into every packet of the traffic that is steered into the SR paths. Since this document aims to define the control plane, it is to be noted that a relevant document for the data plane is [I-D.ietf-ippm-ioam-ipv6-options] for Segment Routing over IPv6 data plane (SRv6).

4.1.1. IOAM Pre-allocated Trace Option Sub-TLV

The IOAM tracing data is expected to be collected at every node that a packet traverses to ensure visibility into the entire path a packet takes within an IOAM domain. The preallocated tracing option will create pre-allocated space for each node to populate its information.

The format of IOAM pre-allocated trace option Sub-TLV is defined as follows:

```
 0                   1                   2                   3
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-------------------------------+-------------------------------+
|          Type=1               |           Length=8            |
+---------------------------------------------------------------+
|       Namespace ID            |            Rsvd1              |
+-------------------------------+-----------------------+-------+
|         IOAM Trace Type                               | Flags | Rsvd2 |
+-------------------------------------------------------+-------+-------+
```

Fig. 4 IOAM Pre-allocated Trace Option Sub-TLV

Where:

Type: 1 (to be assigned by IANA).

Length: 8. It is the total length of the value field not including Type and Length fields.

Namespace ID: A 16-bit identifier of an IOAM-Namespace. The definition is the same as described in section 4.4 of [I-D.ietf-ippm-ioam-data].

IOAM Trace Type: A 24-bit identifier which specifies which data types are used in the node data list. The definition is the same as described in section 4.4 of [I-D.ietf-ippm-ioam-data].

Flags: A 4-bit field. The definition is the same as described in [I-D.ietf-ippm-ioam-flags] and section 4.4 of [I-D.ietf-ippm-ioam-data].

Rsvd1: A 16-bit field reserved for further usage. It MUST be zero and ignored on receipt.

Rsvd2: A 4-bit field reserved for further usage. It MUST be zero and ignored on receipt.

4.1.2. IOAM Incremental Trace Option Sub-TLV

The incremental tracing option contains a variable node data fields where each node allocates and pushes its node data immediately following the option header.

The format of IOAM incremental trace option Sub-TLV is defined as follows:

```
 0                   1                   2                   3
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-------------------------------+-------------------------------+
|          Type=2               |           Length=8            |
+---------------------------------------------------------------+
|       Namespace ID            |            Rsvd1              |
+-------------------------------+-----------------------+-------+
|         IOAM Trace Type                               | Flags | Rsvd2 |
+-----------------------------------------------------+--------+-------+
```

Fig. 5 IOAM Incremental Trace Option Sub-TLV

Where:

Type: 2 (to be assigned by IANA).

Length: 8. It is the total length of the value field not including Type and Length fields.

All the other fields definition is the same as the pre-allocated trace option Sub-TLV in the previous section.

4.1.3. IOAM Directly Export Option Sub-TLV

IOAM directly export option is used as a trigger for IOAM data to be directly exported to a collector without being pushed into in-flight data packets.

The format of IOAM directly export option Sub-TLV is defined as follows:

```
 0                   1                   2                   3
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-------------------------------+-------------------------------+
|           Type=3              |         Length=12             |
+---------------------------------------------------------------+
|        Namespace ID           |            Flags              |
+-------------------------------+---------------+---------------+
|               IOAM Trace Type                 |      Rsvd     |
+-----------------------------------------------+---------------+
|                        Flow ID                                |
+---------------------------------------------------------------+
```

Fig. 6 IOAM Directly Export Option Sub-TLV

Where:

Type: 3 (to be assigned by IANA).

Length: 12. It is the total length of the value field not including Type and Length fields.

Namespace ID: A 16-bit identifier of an IOAM-Namespace. The definition is the same as described in section 4.4 of [I-D.ietf-ippm-ioam-data].

IOAM Trace Type: A 24-bit identifier which specifies which data types are used in the node data list. The definition is the same as described in section 4.4 of [I-D.ietf-ippm-ioam-data].

Flags: A 16-bit field. The definition is the same as described in section 3.2 of [I-D.ietf-ippm-ioam-direct-export].

Flow ID: A 32-bit flow identifier. The definition is the same as described in section 3.2 of [I-D.ietf-ippm-ioam-direct-export].

Rsvd: A 4-bit field reserved for further usage. It MUST be zero and ignored on receipt.

4.1.4. IOAM Edge-to-Edge Option Sub-TLV

The IOAM edge to edge option is to carry data that is added by the IOAM encapsulating node and interpreted by IOAM decapsulating node.

The format of IOAM edge-to-edge option Sub-TLV is defined as follows:

```
 0                   1                   2                   3
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-------------------------------+-------------------------------+
|         Type=4                |          Length=4             |
+---------------------------------------------------------------+
|        Namespace ID           |         IOAM E2E Type         |
+-------------------------------+-------------------------------+
```

Fig. 7 IOAM Edge-to-Edge Option Sub-TLV

Where:

Type: 4 (to be assigned by IANA).

Length: 4. It is the total length of the value field not including Type and Length fields.

Namespace ID: A 16-bit identifier of an IOAM-Namespace. The definition is the same as described in section 4.6 of [I-D.ietf-ippm-ioam-data].

IOAM E2E Type: A 16-bit identifier which specifies which data types are used in the E2E option data. The definition is the same as described in section 4.6 of [I-D.ietf-ippm-ioam-data].

4.2. Enhanced Alternate Marking Sub-TLV

The Alternate Marking [RFC8321]technique is an hybrid performance measurement method, per RFC 7799 [RFC7799] classification of measurement methods. Because this method is based on marking consecutive batches of packets. It can be used to measure packet loss, latency, and jitter on live traffic.

For the SR use case, since this document aims to define the control plane, it is to be noted that a relevant document for the data plane is [I-D.ietf-6man-ipv6-alt-mark] for Segment Routing over IPv6 data plane (SRv6).

The format of Enhanced Alternate Marking (EAM) Sub-TLV is defined as follows:

```
 0                   1                   2                   3
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-------------------------------+-------------------------------+
|            Type=5             |           Length=4            |
+-------------------------------+-------+---------------+-------+
|           FlowMonID                   |     Period    | Flags |
+---------------------------------------+---------------+-------+
```

Fig. 8 Enhanced Alternate Marking Sub-TLV

Where:

Type: 5 (to be assigned by IANA).

Length: 4. It is the total length of the value field not including Type and Length fields.

FlowMonID: A 20-bit identifier to uniquely identify a monitored flow within the measurement domain. The definition is the same as described in section 5.3 of [I-D.ietf-6man-ipv6-alt-mark]. It is to be noted that PCE also needs to maintain the uniqueness of FlowMonID as described in [I-D.ietf-6man-ipv6-alt-mark].

Period: Time interval between two alternate marking period. The unit is second.

Flags: A 4-bits field. Two flags are currently assigned:

H: A flag indicating that the measurement is Hop-By-Hop.

E: A flag indicating that the measurement is End-to-End.

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

5. PCEP Messages

5.1. The PCInitiate Message

A PCInitiate message is a PCEP message sent by a PCE to a PCC to trigger LSP instantiation or deletion RFC 8281 [RFC8281].

For the PCE-initiated LSP with the IFIT feature enabled, IFIT-ATTRIBUTES TLV MUST be included in the LSPA object with the PCInitiate message.

The Routing Backus-Naur Form (RBNF) definition of the PCInitiate message RFC 8281 [RFC8281] is unchanged by this document.

5.2. The PCUpd Message

A PCUpd message is a PCEP message sent by a PCE to a PCC to update the LSP parameters RFC 8231 [RFC8231].

For PCE-initiated LSPs with the IFIT feature enabled, the IFIT-ATTRIBUTES TLV MUST be included in the LSPA object with the PCUpd message. The PCE can send this TLV to direct the PCC to change the IFIT parameters.

The RBNF definition of the PCUpd message RFC 8231 [RFC8231] is unchanged by this document.

5.3. The PCRpt Message

The PCRpt message RFC 8231 [RFC8231] is a PCEP message sent by a PCC to a PCE to report the status of one or more LSPs.

For PCE-initiated LSPs RFC 8281 [RFC8281], the PCC creates the LSP using the attributes communicated by the PCE and the local values for the unspecified parameters. After the successful instantiation of the LSP, the PCC automatically delegates the LSP to the PCE and generates a PCRpt message to provide the status report for the LSP.

The RBNF definition of the PCRpt message RFC 8231 [RFC8231] is unchanged by this document.

For both PCE-initiated and PCC-initiated LSPs, when the LSP is instantiated the IFIT methods are applied as specified for the corresponding data plane. [I-D.ietf-ippm-ioam-ipv6-options] and [I-D.ietf-6man-ipv6-alt-mark] are the relevant documents for Segment Routing over IPv6 data plane (SRv6).

6. Example of application to SR Policy

A PCC or PCE sets the IFIT-CAPABILITY TLV in the Open message during the PCEP initialization phase to indicate that it supports the IFIT procedures.

[I-D.ietf-pce-segment-routing-policy-cp] defines the PCEP extension to support Segment Routing Policy Candidate Paths and in this regard the SRPAG Association object is introduced.

The Examples of PCC Initiated SR Policy with single or multiple candidate-paths and PCE Initiated SR Policy with single or multiple candidate-paths are reported in [I-D.ietf-pce-segment-routing-policy-cp].

In case of PCC Initiated SR Policy, PCC sends PCReq message to the PCE, encoding the SRPAG ASSOCIATION object and IFIT-ATTRIBUTES TLV via the LSPA object. This is valid for both single and multiple candidate-paths. Finally PCE returns the path in PCRep message, and echoes back the SRPAG object that were used in the computation and IFIT LSPA TLVs too. Additionally, PCC sends PCRpt message to the PCE, including the LSP object and the SRPAG ASSOCIATION object and IFIT-ATTRIBUTES TLV via the LSPA object. Then PCE computes path and finally PCE updates the SR policy candidate path's ERO using PCUpd message considering the IFIT LSPA TLVs too.

In case of PCE Initiated SR Policy, PCE sends PCInitiate message, containing the SRPAG Association object and IFIT-ATTRIBUTES TLV via the LSPA object. This is valid for both single and multiple candidate-paths. Then PCC uses the color, endpoint and preference from the SRPAG object to create a new candidate path considering the IFIT LSPA TLVs too. Finally PCC sends a PCRpt message back to the PCE to report the newly created Candidate Path. The PCRpt message contains the SRPAG Association object and IFIT-ATTRIBUTES information.

The procedure of enabling/disabling IFIT is simple, indeed the PCE can update the IFIT-ATTRIBUTES of the LSP by sending subsequent Path Computation Update Request (PCUpd) messages. PCE can update the IFIT-ATTRIBUTES of the LSP by sending Path Computation State Report (PCRpt) messages.

7. IANA Considerations

This document defines the new IFIT-CAPABILITY TLV and IFIT-ATTRIBUTES TLV.

7.1. PCEP TLV Type Indicators

IANA is requested to make the assignment from the "PCEP TLV Type Indicators" subregistry of the "Path Computation Element Protocol (PCEP) Numbers" registry as follows:

Value	Description	Reference
TBD1	IFIT-CAPABILITY TLV	This document
TBD2	IFIT-ATTRIBUTES TLV	This document

7.2. IFIT-CAPABILITY TLV Flags field

This document specifies the IFIT-CAPABILITY TLV 32-bits Flags field. IANA is requested to create a registry to manage the value of the IFIT-CAPABILITY TLV's Flags field within the "Path Computation Element Protocol (PCEP) Numbers" registry.

New values are to be assigned by Standards Action RFC 8126 [RFC8126]. Each bit should be tracked with the following qualities:

* Bit number (count from 0 as the most significant bit)

* Flag Name

* Reference

IANA is requested to set 5 new bits in the IFIT-CAPABILITY TLV Flags Field registry, as follows:

Bit no.	Flag Name	Reference
0-26	Unassigned	This document
27	P: IOAM Pre-allocated Trace Option flag	This document
28	I: IOAM Incremental Trace Option flag	This document
29	D: IOAM Directly Export Option flag	This document
30	E: IOAM Edge-to-Edge Option	This document
31	M: Alternate Marking Flag	This document

7.3. IFIT-ATTRIBUTES Sub-TLV

This document also specifies the IFIT-ATTRIBUTES sub-TLVs. IANA is requested to create an "IFIT-ATTRIBUTES Sub-TLV Types" subregistry within the "Path Computation Element Protocol (PCEP) Numbers" registry.

IANA is requested to set the Registration Procedure for this registry to read as follows:

Range	Registration Procedure
0-65503	IETF Review
65504-65535	Experimental Use

This document defines the following types:

Type	Description	Reference
0	Reserved	This document
1	IOAM Pre-allocated Trace Option	This document
2	IOAM Incremental Trace Option	This document
3	IOAM Directly Export Option	This document
4	IOAM Edge-to-Edge Option	This document
5	Enhanced Alternate Marking	This document
6-65503	Unassigned	This document
65504-65535	Experimental Use	This document

7.4. Enhanced Alternate Marking Sub-TLV Flags field

This document specifies the Enhanced Alternate Marking Sub-TLV 4-bits Flags field. IANA is requested to create a registry to manage the value of the Enhanced Alternate Marking Sub-TLV's Flags field within the "Path Computation Element Protocol (PCEP) Numbers" registry.

New values are to be assigned by Standards Action RFC 8126 [RFC8126]. Each bit should be tracked with the following qualities:

* Bit number (count from 0 as the most significant bit)

* Flag Name

* Reference

IANA is requested to set 2 new bits in the IFIT-CAPABILITY TLV Flags Field registry, as follows:

Bit no.	Flag Name	Reference
3	H: Hop-By-Hop flag	This document
2	E: End-to-End flag	This document
0-1	Unassigned	

7.5. PCEP Error Codes

This document defines a new Error-value for PCErr message of Error-Type 19 (Invalid Operation). IANA is requested to allocate a new Error-value within the "PCEP-ERROR Object Error Types and Values" subregistry of the "Path Computation Element Protocol (PCEP) Numbers" registry as follows:

Error-Type	Meaning	Error-value	Reference
19	Invalid Operation	TBD3: IFIT capability not advertised	This document

8. Security Considerations

This document defines the new IFIT-CAPABILITY TLV and IFIT Attributes TLVs, which do not add any substantial new security concerns beyond those already discussed in RFC 8231 [RFC8231] and RFC 8281 [RFC8281] for stateful PCE operations. As per RFC 8231 [RFC8231], it is RECOMMENDED that these PCEP extensions only be activated on authenticated and encrypted sessions across PCEs and PCCs belonging to the same administrative authority, using Transport Layer Security (TLS) RFC 8253 [RFC8253], as per the recommendations and best current practices in BCP 195 RFC 7525 [RFC7525] (unless explicitly set aside in RFC 8253 [RFC8253]).

Implementation of IFIT methods (IOAM and Alternate Marking) are mindful of security and privacy concerns, as explained in [I-D.ietf-ippm-ioam-data] and RFC 8321 [RFC8321]. Anyway incorrect IFIT parameters in the IFIT-ATTRIBUTES sub-TLVs SHOULD NOT have an adverse effect on the LSP as well as on the network, since it affects only the operation of the telemetry methodology.

IFIT data MUST be propagated in a limited domain in order to avoid malicious attacks and solutions to ensure this requirement are respectively discussed in [I-D.ietf-ippm-ioam-data] and [I-D.ietf-6man-ipv6-alt-mark].

IFIT methods (IOAM and Alternate Marking) are applied within a controlled domain where the network nodes are locally administered. A limited administrative domain provides the network administrator with the means to select, monitor and control the access to the network, making it a trusted domain also for the PCEP extensions defined in this document.

9. Contributors

The following people provided relevant contributions to this document:

Huanan Chen, independent, -

Dhruv Doody, Huawei Technologies, dhruv.ietf@gmail.com

10. Acknowledgements

The authors of this document would like to thank Huaimo Chen for the comments and review of this document.

11. References

11.1. Normative References

[I-D.ietf-6man-ipv6-alt-mark]
Fioccola, G., Zhou, T., Cociglio, M., Qin, F., and R. Pang, "IPv6 Application of the Alternate Marking Method", draft-ietf-6man-ipv6-alt-mark-12 (work in progress), October 2021.

[I-D.ietf-ippm-ioam-data]
Brockners, F., Bhandari, S., and T. Mizrahi, "Data Fields for In-situ OAM", draft-ietf-ippm-ioam-data-17 (work in progress), December 2021.

[I-D.ietf-ippm-ioam-direct-export]
Song, H., Gafni, B., Zhou, T., Li, Z., Brockners, F., Bhandari, S., Sivakolundu, R., and T. Mizrahi, "In-situ OAM Direct Exporting", draft-ietf-ippm-ioam-direct-export-07 (work in progress), October 2021.

[I-D.ietf-ippm-ioam-flags]
Mizrahi, T., Brockners, F., Bhandari, S., Sivakolundu, R., Pignataro, C., Kfir, A., Gafni, B., Spiegel, M., and J. Lemon, "In-situ OAM Loopback and Active Flags", draft-ietf-ippm-ioam-flags-07 (work in progress), October 2021.

[I-D.ietf-ippm-ioam-ipv6-options]
Bhandari, S. and F. Brockners, "In-situ OAM IPv6 Options", draft-ietf-ippm-ioam-ipv6-options-06 (work in progress), July 2021.

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <https://www.rfc-editor.org/info/rfc2119>.

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

[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>.

[RFC7799] Morton, A., "Active and Passive Metrics and Methods (with Hybrid Types In-Between)", RFC 7799, DOI 10.17487/RFC7799, May 2016, <https://www.rfc-editor.org/info/rfc7799>.

[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>.

[RFC8174] Leiba, B., "Ambiguity of Uppercase vs Lowercase in RFC 2119 Key Words", BCP 14, RFC 8174, DOI 10.17487/RFC8174, May 2017, <https://www.rfc-editor.org/info/rfc8174>.

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

[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>.

[RFC8321] Fioccola, G., Ed., Capello, A., Cociglio, M., Castaldelli, L., Chen, M., Zheng, L., Mirsky, G., and T. Mizrahi, "Alternate-Marking Method for Passive and Hybrid Performance Monitoring", RFC 8321, DOI 10.17487/RFC8321, January 2018, <https://www.rfc-editor.org/info/rfc8321>.

[RFC8664] Sivabalan, S., Filsfils, C., Tantsura, J., Henderickx, W., and J. Hardwick, "Path Computation Element Communication Protocol (PCEP) Extensions for Segment Routing", RFC 8664, DOI 10.17487/RFC8664, December 2019, <https://www.rfc-editor.org/info/rfc8664>.

[RFC8799] Carpenter, B. and B. Liu, "Limited Domains and Internet Protocols", RFC 8799, DOI 10.17487/RFC8799, July 2020, <https://www.rfc-editor.org/info/rfc8799>.

11.2. Informative References

[I-D.ietf-pce-segment-routing-ipv6]
Li, C., Negi, M., Sivabalan, S., Koldychev, M., Kaladharan, P., and Y. Zhu, "PCEP Extensions for Segment Routing leveraging the IPv6 data plane", draft-ietf-pce-segment-routing-ipv6-11 (work in progress), January 2022.

[I-D.ietf-pce-segment-routing-policy-cp]
Koldychev, M., Sivabalan, S., Barth, C., Peng, S., and H. Bidgoli, "PCEP extension to support Segment Routing Policy Candidate Paths", draft-ietf-pce-segment-routing-policy-cp-06 (work in progress), October 2021.

[I-D.ietf-spring-segment-routing-policy]
Filsfils, C., Talaulikar, K., Voyer, D., Bogdanov, A., and P. Mattes, "Segment Routing Policy Architecture", draft-ietf-spring-segment-routing-policy-16 (work in progress), January 2022.

[I-D.koldychev-pce-multipath]
Koldychev, M., Sivabalan, S., Saad, T., Beeram, V. P., Bidgoli, H., Yadav, B., and S. Peng, "PCEP Extensions for Signaling Multipath Information", draft-koldychev-pce-multipath-05 (work in progress), February 2021.

[I-D.qin-idr-sr-policy-ifit]
Qin, F., Yuan, H., Zhou, T., Fioccola, G., and Y. Wang, "BGP SR Policy Extensions to Enable IFIT", draft-qin-idr-sr-policy-ifit-04 (work in progress), October 2020.

Authors' Addresses

Hang Yuan
UnionPay
1899 Gu-Tang Rd., Pudong
Shanghai
China

Email: yuanhang@unionpay.com

Tianran Zhou
Huawei
156 Beiqing Rd., Haidian District
Beijing
China

Email: zhoutianran@huawei.com

Weidong Li
Huawei
156 Beiqing Rd., Haidian District
Beijing
China

Email: poly.li@huawei.com

Giuseppe Fioccola
Huawei
Riesstrasse, 25
Munich
Germany

Email: giuseppe.fioccola@huawei.com

Yali Wang
Huawei
156 Beiqing Rd., Haidian District
Beijing
China

Email: wangyali11@huawei.com

Network Working Group
Internet-Draft
Updates: 5440 (if approved)
Intended status: Standards Track
Expires: 3 August 2022

A. Stone
M. Aissaoui
Nokia
S. Sidor
Cisco Systems, Inc.
S. Sivabalan
Ciena Coroporation

30 January 2022

Local Protection Enforcement in PCEP
draft-ietf-pce-local-protection-enforcement-04

Abstract

This document updates [RFC5440] to clarify usage of the local protection desired bit signalled in Path Computation Element Protocol (PCEP). This document also introduces a new flag for signalling protection strictness in PCEP.

Status of This Memo

Copyright Notice

Table of Contents

1. Introduction

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

PCEP [RFC5440] utilizes flags, values and concepts previously defined in RSVP-TE Extensions [RFC3209] and Fast Reroute Extensions to RSVP-TE [RFC4090]. One such concept in PCEP is the 'Local Protection Desired' (L-flag in the LSPA Object in [RFC5440]), which was originally defined in the SESSION-ATTRIBUTE Object in RFC3209. In RSVP, this flag signals to downstream routers that local protection is desired, which indicates to transit routers that they may use a local repair mechanism. The headend router calculating the path does not know whether a downstream router will or will not protect a hop during it's calculation. Therefore, a local protection desired does

not require the transit router to satisfy protection in order to establish the RSVP signalled path. This flag is signalled in PCEP as an attribute of the LSP via the LSP Attributes object.

PCEP Extensions for Segment Routing ([RFC8664]) extends support in PCEP for Segment Routed LSPs (SR-LSPs) as defined in the Segment Routing Architecture [RFC8402]. As per the Segment Routing Architecture, Adjacency Segment Identifiers(Adj-SID) may be eligible for protection (using IPFRR or MPLS-FRR). The protection eligibility is advertised into IGP ([RFC8665] and [RFC8667]) as the B-Flag part of the Adjacency SID sub-tlv and can be discovered by a PCE via BGP-LS [RFC7752] using the BGP-LS Segment Routing Extensions ([RFC9085]). An Adjacency SID may or may not have protection eligibility and for a given adjacency between two routers there may be multiple Adjacency SIDs, some of which are protected and some which are not.

A Segment Routed path calculated by PCE may contain various types of segments, as defined in [RFC8402] such as Adjacency, Node or Binding. The protection eligibility for Adjacency SIDs can be discovered by PCE, so therefore the PCE can take the protection eligibility into consideration as a path constraint. If a path is calculated to include other segment identifiers which are not applicable to having their protection state advertised, as they may only be locally significant for each router processing the SID such as Node SIDs, it may not be possible for PCE to include the protection constraint as part of the path calculation.

It is desirable for an operator to define the enforcement, or strictness of the protection requirement when it can be applied.

This document updates [RFC5440] by further describing the behaviour with Local Protection Desired Flag (L-Flag) and extends on it with the introduction of Enforcement Flag (E-Flag).

2. Requirements Language

In this document, the key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" are to be interpreted as described in BCP 14, [RFC2119].

3. Terminology

This document uses the following terminology:

PROTECTION MANDATORY: path MUST have protection eligibility on all links.

UNPROTECTED MANDATORY: path MUST NOT have protection eligibility on all links.

PROTECTION PREFERRED: path SHOULD have protection eligibility on all links but MAY contain links which do not have protection eligibility.

UNPROTECTED PREFERRED: path SHOULD NOT have protection eligibility on all links but MAY contain links which have protection eligibility.

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

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

PCEP: Path Computation Element Protocol.

4. Motivation

4.1. Implementation differences

As defined in [RFC5440] the mechanism to signal protection enforcement in PCEP is with the previously mentioned L-flag defined in the LSPA Object. The name of the flag uses the term "Desired", which by definition means "strongly wished for or intended" and the use case originated from the RSVP. For RSVP signalled paths, local protection is not within control of the PCE. However, [RFC5440] does state "When set, this means that the computed path must include links protected with Fast Reroute as defined in [RFC4090]." Implementations of [RFC5440] have either interpreted the L-Flag as PROTECTION MANDATORY or PROTECTION PREFERRED, leading to operational differences.

4.2. SLA Enforcement

The boolean bit flag is unable to distinguish between the different options of PROTECTION MANDATORY, UNPROTECTED MANDATORY, PROTECTION PREFERRED and UNPROTECTED PREFERRED. The selection of the options are typically dependent on the service level agreement the operator wishes to impose on the LSP. When enforcement is used, the resulting shortest path calculation is impacted.

For example, PROTECTION MANDATORY is for use cases where an operator may need the LSP to follow a path which has local protection provided along the full path, ensuring that if there is anywhere along the path that traffic will be fast re-routed at the point of failure.

For another example, UNPROTECTED MANDATORY is when an operator may intentionally prefer an LSP to not be locally protected, and thus would rather local failures to cause the LSP to go down and/or rely on other protection mechanisms such as a secondary diverse path.

There are also use cases where there is simply no requirement to enforce protection or no protection along a path. This can be considered as "do not care to enforce". This is a relaxation of the protection constraint. The path calculation is permitted the use of any SID which is available along the calculated path. The SID backup availability does not impact the shortest path computation. Since links may have both protected and unprotected SIDs available, the option PROTECTION PREFERRED or UNPROTECTED PREFERRED is used to instruction PCE a preference on which SID to select, as the behaviour of the LSP would differ during a local failure depending on which SID is selected.

5. Protection Enforcement Flag (E-Flag)

Section 7.11 in Path Computation Element Protocol [RFC5440] describes the encoding of the Local Protection Desired (L-Flag). A new flag is proposed in this document in the LSP Attributes Object which extends the L-Flag to identify the protection enforcement.

Bit 6 has been early allocated by IANA as the Protection Enforcement flag.

Codespace of the Flag field (LSPA Object)

Bit	Description	Reference
7	Local Protection Desired	RFC5440
6	Local Protection Enforcement	This I-D

The format of the LSPA Object as defined in [RFC5440] is:

```
 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
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                       Exclude-any                             |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                       Include-any                             |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                       Include-all                             |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|  Setup Prio   |  Holding Prio |     Flags |E|L|   Reserved    |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                                                               |
//                     Optional TLVs                           //
|                                                               |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Flags (8 bits)

* L flag: As defined in [RFC5440] and further updated by this document. When set, protection is desired. When not set, protection is not desired. The enforcement of the protection is identified via the E-Flag.

* E flag (Protection Enforcement): When set, the value of the L-Flag MUST be treated as a MUST constraint where applicable, when protection state of a SID is known. When E flag is not set, the value of the L-Flag MUST be treated as a MAY constraint.

When L-flag is set and E-flag is set then PCE MUST consider the protection eligibility as PROTECTION MANDATORY constraint.

When L-flag is set and E-flag is not set then PCE MUST consider the protection eligibility as PROTECTION PREFERRED constraint.

When L-flag is not set and E-flag is not set then PCE SHOULD consider the protection eligibility as UNPROTECTED PREFERRED but MAY consider protection eligibility as UNPROTECTED MANDATORY constraint.

When L-flag is not set and E-flag is set then PCE MUST consider the protection eligibility as UNPROTECTED MANDATORY constraint.

UNPROTECTED PREFERRED and PROTECTED PREFERRED may seem similar but they indicate the preference of selection of a SID if PCE has an option of either protected or unprotected available on a link. When presented with either option, PCE SHOULD select the SID which has a protection state matching the state of the L-Flag.

The protection enforcement constraint can only be applied to resource selection in which the protection state is known to PCE. A PCE calculating a path that includes resources which does not support the protection state being known to PCE (such as Node SID), then the protection state MAY ignore the protection enforcement constraint.

5.1. Backwards Compatibility

Considerations in the message passing between PCC and PCE for the E-Flag bit which are not supported by the entity are outlined in this section, with requirements for PCE and PCC implementing this document described at the end.

For a PCC or PCE which does not yet support this document, the E-flag bit is ignored and set to zero in PCRpt and/or PCUpd as per [RFC5440] for PCC-initiated or as per ([RFC8281]) for PCE-initiated LSPs. It's important to note that [RFC8231] and [RFC8281] permit LSP Attribute Object to be included in PCUpd messages for PCC-initiated and PCE-initiated LSPs. For PCC-initiated LSPs, PCUpd E-Flag (and L-Flag) are an echo from the previous PCRpt however the bit value is ignored on PCE from the previous PCRpt, therefore the E-Flag value set in the PCUpd is zero. A PCE which does not support this document sends PCUpd messages with the E-Flag unset for PCC-initated LSPs even if set in the prior PCReq or PCRpt. A PCC which does not support this document sends PCRpt messages with the E-Flag unset for PCE-initiated LSPs even if set in the prior PCInitiate or PCUpd.

For a PCC which does support this document, it MAY set E-Flag bit depending on local configuration. If communicating with a PCE which does not yet support this document, the PCE follows the behaviour specified in [RFC5440] and will ignore the E-Flag bit thus it will not compute a path respecting the enforcement constraint.

For PCC-initiated LSPs, PCC SHOULD ignore the E-Flag value received from PCE in a PCUpd message.

For PCE-initiated LSPs, PCC MAY process the E-Flag value received from PCE in a PCUpd message. PCE SHOULD ignore the E-Flag value received from PCC in a PCRpt message.

6. Implementation Status

[Note to the RFC Editor - remove this section before publication, as well as remove the reference to RFC 7942.]

This section records the status of known implementations of the protocol defined by this specification at the time of posting of this Internet-Draft, and is based on a proposal described in [RFC7942].

The description of implementations in this section is intended to assist the IETF in its decision processes in progressing drafts to RFCs. Please note that the listing of any individual implementation here does not imply endorsement by the IETF. Furthermore, no effort has been spent to verify the information presented here that was supplied by IETF contributors. This is not intended as, and must not be construed to be, a catalogue of available implementations or their features. Readers are advised to note that other implementations may exist.

According to [RFC7942], "this will allow reviewers and working groups to assign due consideration to documents that have the benefit of running code, which may serve as evidence of valuable experimentation and feedback that have made the implemented protocols more mature. It is up to the individual working groups to use this information as they see fit".

6.1. Nokia Implementation

* Organization: Nokia

* Implementation: NSP PCE and SROS PCC.

* Description: Implementation for calculation and conveying intention described in this document

* Maturity Level: Demo

* Coverage: Full

* Contact: andrew.stone@nokia.com

6.2. Cisco Implementation

* Organization: Cisco Systems, Inc.

* Implementation: IOS-XR PCE and PCC.

* Description: Implementation for calculation and conveying intention described in this document

* Maturity Level: Demo

* Coverage: Full

* Contact: ssidor@cisco.com

7. Security Considerations

This document clarifies the behaviour of an existing flag and introduces a new flag to provide further control of that existing behaviour. The introduction of this new flag and behaviour clarification does not create any new sensitive information. No additional security measure is required.

Securing the PCEP session using Transport Layer Security (TLS) [RFC8253], as per the recommendations and best current practices in [RFC7525] is RECOMMENDED.

8. IANA Considerations

8.1. LSPA Object

This document defines a new bit value in the sub-registry "LSPA Object Flag Field" in the "Path Computation Element Protocol (PCEP) Numbers" registry. IANA is requested to confirm the early-allocated codepoint.

Bit	Name	Reference
6	Protection Enforcement	This I-D

9. References

9.1. Normative References

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <https://www.rfc-editor.org/info/rfc2119>.

[RFC4655] Farrel, A., Vasseur, J.-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>.

[RFC8402] Filsfils, C., Ed., Previdi, S., Ed., Ginsberg, L., Decraene, B., Litkowski, S., and R. Shakir, "Segment Routing Architecture", RFC 8402, DOI 10.17487/RFC8402, July 2018, <https://www.rfc-editor.org/info/rfc8402>.

[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>.

[RFC3209] Awduche, D., Berger, L., Gan, D., Li, T., Srinivasan, V., and G. Swallow, "RSVP-TE: Extensions to RSVP for LSP Tunnels", RFC 3209, DOI 10.17487/RFC3209, December 2001, <https://www.rfc-editor.org/info/rfc3209>.

[RFC4090] Pan, P., Ed., Swallow, G., Ed., and A. Atlas, Ed., "Fast Reroute Extensions to RSVP-TE for LSP Tunnels", RFC 4090, DOI 10.17487/RFC4090, May 2005, <https://www.rfc-editor.org/info/rfc4090>.

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

[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>.

[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>.

[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>.

[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>.

9.2. Informative References

[RFC7942] Sheffer, Y. and A. Farrel, "Improving Awareness of Running Code: The Implementation Status Section", BCP 205, RFC 7942, DOI 10.17487/RFC7942, July 2016, <https://www.rfc-editor.org/info/rfc7942>.

[RFC8664] Sivabalan, S., Filsfils, C., Tantsura, J., Henderickx, W., and J. Hardwick, "Path Computation Element Communication Protocol (PCEP) Extensions for Segment Routing", RFC 8664, DOI 10.17487/RFC8664, December 2019, <https://www.rfc-editor.org/info/rfc8664>.

[RFC8665] Psenak, P., Ed., Previdi, S., Ed., Filsfils, C., Gredler, H., Shakir, R., Henderickx, W., and J. Tantsura, "OSPF Extensions for Segment Routing", RFC 8665, DOI 10.17487/RFC8665, December 2019, <https://www.rfc-editor.org/info/rfc8665>.

[RFC8667] Previdi, S., Ed., Ginsberg, L., Ed., Filsfils, C., Bashandy, A., Gredler, H., and B. Decraene, "IS-IS Extensions for Segment Routing", RFC 8667, DOI 10.17487/RFC8667, December 2019, <https://www.rfc-editor.org/info/rfc8667>.

[RFC9085] Previdi, S., Talaulikar, K., Ed., Filsfils, C., Gredler, H., and M. Chen, "Border Gateway Protocol - Link State (BGP-LS) Extensions for Segment Routing", RFC 9085, DOI 10.17487/RFC9085, August 2021, <https://www.rfc-editor.org/info/rfc9085>.

Acknowledgements

Thanks to Dhruv Dhody and Mike Koldychev for reviewing, commenting and discussions on this document.

Authors' Addresses

Andrew Stone
Nokia
600 March Road
Kanata Ontario K2K 2T6
Canada

Email: andrew.stone@nokia.com

Mustapha Aissaoui
Nokia
600 March Road

Kanata Ontario K2K 2T6
Canada

Email: mustapha.aissaoui@nokia.com

Samuel Sidor
Cisco Systems, Inc.
Eurovea Central 3.
Pribinova 10
811 09 Bratislava
Slovakia

Email: ssidor@cisco.com

Siva Sivabalan
Ciena Coroporation
385 Terry Fox Drive
Kanata Ontario K2K 0L1
Canada

Email: ssivabal@ciena.com

Network Working Group
Internet-Draft
Updates: 5440 (if approved)
Intended status: Standards Track
Expires: 25 December 2023

A. Stone
M. Aissaoui
Nokia
S. Sidor
Cisco Systems, Inc.
S. Sivabalan
Ciena Coroporation
23 June 2023

Local Protection Enforcement in PCEP
draft-ietf-pce-local-protection-enforcement-11

Abstract

This document updates RFC5440 to clarify usage of the local protection desired bit signalled in the Path Computation Element Protocol (PCEP). This document also introduces a new flag for signalling protection strictness in PCEP.

Status of This Memo

Copyright Notice

Table of Contents

1. Introduction

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

PCEP [RFC5440] utilizes flags, values and concepts previously defined in RSVP-TE Extensions [RFC3209] and Fast Reroute Extensions to RSVP-TE [RFC4090]. One such concept in PCEP is the 'Local Protection Desired' (L flag in the LSPA Object in [RFC5440]), which was originally defined in the SESSION-ATTRIBUTE Object in RFC3209. In RSVP, this flag signals to downstream routers that they may use a local repair mechanism. The headend router calculating the path does not know whether a downstream router will or will not protect a hop during its calculation. Therefore, a local protection desired does not require the transit router to satisfy protection in order to establish the RSVP signalled path. This flag is signalled in PCEP as an attribute of the LSP via the LSP Attributes object.

PCEP Extensions for Segment Routing ([RFC8664]) extends support in PCEP for Segment Routed paths. The path list is encoded with Segment Identifiers, each of which might offer local protection. The PCE may discover the protection eligibility for a Segment Identifier (SID) via BGP-LS [RFC9085] and take the protection into consideration as a path constraint.

It is desirable for an operator to be able to define the enforcement, or strictness of the protection requirement.

This document updates [RFC5440] by further describing the behaviour with the Local Protection Desired Flag (L flag) and extends on it with the introduction of the Enforcement Flag (E flag).

The document contains reference notes for Segment Routing, however the content described is path setup type and data plane technology agnostic.

2. Requirements Language

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

3. Terminology

This document uses the following terminology:

PROTECTION MANDATORY: The Path MUST have protection eligibility on all links.

UNPROTECTED MANDATORY: The Path MUST NOT have protection eligibility on all links.

PROTECTION PREFERRED: The Path should have protection eligibility on all links but might contain links which do not have protection eligibility.

UNPROTECTED PREFERRED: The Path should not have protection eligibility on all links but might contain links which have protection eligibility.

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

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

PCEP: Path Computation Element Protocol.

LSPA: LSP Attributes Object in PCEP, defined in RFC5440

4. Motivation

4.1. Implementation differences

As defined in [RFC5440] the mechanism to signal protection enforcement in PCEP is the previously mentioned L flag defined in the LSPA Object. The name of the flag uses the term "Desired", which by definition means "strongly wished for or intended" and the use case originated from the RSVP. For RSVP signalled paths, local protection is not within control of the PCE. However, [RFC5440] does state "When set, this means that the computed path must include links protected with Fast Reroute as defined in [RFC4090]." Implementations of [RFC5440] have either interpreted the L flag as PROTECTION MANDATORY or PROTECTION PREFERRED, leading to operational differences.

4.2. SLA Enforcement

The boolean bit L flag is unable to distinguish between the different options of PROTECTION MANDATORY, UNPROTECTED MANDATORY, PROTECTION PREFERRED and UNPROTECTED PREFERRED. Selecting one of the options is typically dependent on the service level agreement the operator wishes to impose on the LSP. A network may be providing transit to multiple service agreement definitions against the same base topology network, whose behavior could vary, such as wanting local protection to be invoked on some LSPs and not wanting local protection on others. When enforcement is used, the resulting shortest path calculation is impacted.

For example, PROTECTION MANDATORY is for use cases where an operator may need the LSP to follow a path which has local protection provided along the full path, ensuring that if there is a failure anywhere along the path that traffic will be fast re-routed at the point.

For example, UNPROTECTED MANDATORY is when an operator may intentionally prefer an LSP to not be locally protected, and thus would rather local failures cause the LSP to go down. An example scenario is one where an LSP is protected with path protection via a secondary diverse LSP. Each LSP is traffic engineered to follow specific traffic engineered criteria computed by the PCE to satisfy SLA. Upon a failure, if local protection is invoked on the active LSP traffic, the traffic may temporarily traverse links which violate the TE requirements and could negatively impact the resources being traversed (e.g., insufficient bandwidth). In addition, depending on the network topological scenario, it may be not feasible for the PCE to reroute the LSP while respecting the TE requirements which include path diversity, resulting in the LSP being torn down and switched to the protected path anyways. In such scenarios its desirable for the LSP to be simply torn down immediately and not re-routed through local protection, so that traffic may be forwarded through an already established traffic-engineered secondary path.

Both UNPROTECTED PREFERRED and PROTECTED PREFERRED options provide a relaxation of the protection constraint. These options can be used when an operator does not require protection enforcement. Regardless of the option selected, the protection status of a resource does not influence whether the link must be pruned during a path calculation. Furthermore, the selection of either option indicates a priority selection to PCE when there is an option to choose a protected or unprotected instruction associated with a resource, ensuring consistent PCE behavior across different implementations.

When used with Segment Routing, an adjacency may have both a protected SID and an unprotected SID. If the UNPROTECTED PREFERRED option is selected, PCE chooses the unprotected SID. Alternatively, if the PROTECTED PREFERRED option is selected, PCE chooses the protected SID

5. Protection Enforcement Flag (E flag)

Section 7.11 in Path Computation Element Protocol [RFC5440] describes the encoding of the Local Protection Desired (L flag). A Protection Enforcement flag "E" is specified below, extending the L flag.

[RFC Editor Note: The text below assumes the E bit remains the early allocation value 6. Please adjust if this changes and remove this note before publication.]

Codespace of the Flag field (LSPA Object)

Bit	Description	Reference
7	Local Protection Desired	RFC5440
6	Local Protection Enforcement	This document

The format of the LSPA Object as defined in [RFC5440] is:

```
 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
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                        Exclude-any                            |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                        Include-any                            |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                        Include-all                            |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|  Setup Prio   |  Holding Prio |     Flags |E|L|   Reserved    |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                                                               |
//                      Optional TLVs                          //
|                                                               |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Flags (8 bits)

* L Flag: As defined in [RFC5440] and further updated by this document. When set to 1, protection is desired. When set to 0, protection is not desired. The enforcement of the protection is identified via the E flag.

* E Flag (Protection Enforcement): This flag controls the strictness in which the PCE must apply the L flag. When set to 1, the value of the L flag needs to be respected during resource selection by the PCE. When E flag is set to 0, an attempt to respect the value of the L flag is made; however, the PCE could relax or ignore the L flag when computing a path. The statements below indicate preference when the E flag is set to 0 in combination with the L flag value.

When both the L flag and E flag are set to 1, then the PCE MUST consider the protection eligibility as a PROTECTION MANDATORY constraint.

When the L flag is set to 1 and the E flag is set to 0, then the PCE MUST consider the protection eligibility as a PROTECTION PREFERRED constraint.

When both L flag and E flag are set to 0, then the PCE SHOULD consider the protection eligibility as an UNPROTECTED PREFERRED constraint but MAY consider protection eligibility as an UNPROTECTED MANDATORY constraint. An example of when the latter behavior might be chosen is if the PCE has some means (outside the scope of this document) to detect that it is interacting with a legacy PCC that expects the legacy behavior.

When L flag is set to 0 and E flag is set to 1, then the PCE MUST consider the protection eligibility as an UNPROTECTED MANDATORY constraint.

If a PCE is unable to infer the protection status of a resource, the PCE MAY use local policy to define protected status assumptions. When computing a Segment Routed path, It is RECOMMENDED that a PCE assume a Node SID is protected. It is also RECOMMENDED that a PCE assume an Adjacency SID is protected if the backup flag advertised with the Adjacency SID is set.

5.1. Backwards Compatibility

Considerations in the message passing between the PCC and the PCE for the E flag bit which are not supported by the entity are outlined in this section, with requirements for the PCE and the PCC implementing this document described at the end.

For a PCC or PCE which does not yet support this document, the E flag is ignored and set to zero in PCRpt and/or PCUpd as per [RFC5440] for PCC-initiated or as per [RFC8281] for PCE-initiated LSPs. It is important to note that [RFC8231] and [RFC8281] permit the LSP Attribute Object to be included in PCUpd messages for PCC-initiated and PCE-initiated LSPs.

For PCC-initiated LSPs, PCUpd E flag (and L flag) is an echo from the previous PCRpt however the bit value is ignored on the PCE from the previous PCRpt, therefore the E flag value set in the PCUpd is zero. A PCE which does not support this document sends PCUpd messages with the E flag set to 0 for PCC-initated LSPs even if set to 1 in the prior PCReq or PCRpt.

A PCC which does not support this document sends PCRpt messages with the E flag set to 0 for PCE-initiated LSPs even if set to 1 in the prior PCInitiate or PCUpd.

For a PCC which does support this document, it MAY set the E flag to 1 depending on local configuration. If communicating with a PCE which does not yet support this document, the PCE follows the behaviour specified in [RFC5440] and will ignore the E flag. Thus, a computed path might not respect the enforcement constraint.

For PCC-initiated LSPs, the PCC SHOULD ignore the E flag value received from the PCE in a PCUpd message as it may be communicating with a PCE which does not support this document.

For PCE-initiated LSPs, the PCC MAY process the E flag value received from the PCE in a PCUpd message. The PCE SHOULD ignore the E flag value received from the PCC in a PCRpt message as it may be communicating with a PCC which does not support this document.

6. Implementation Status

[Note to the RFC Editor - remove this section before publication, as well as remove the reference to RFC 7942.]

This section records the status of known implementations of the protocol defined by this specification at the time of posting of this Internet-Draft, and is based on a proposal described in [RFC7942]. The description of implementations in this section is intended to assist the IETF in its decision processes in progressing drafts to RFCs. Please note that the listing of any individual implementation here does not imply endorsement by the IETF. Furthermore, no effort has been spent to verify the information presented here that was supplied by IETF contributors. This is not intended as, and must not be construed to be, a catalogue of available implementations or their features. Readers are advised to note that other implementations may exist.

According to [RFC7942], "this will allow reviewers and working groups to assign due consideration to documents that have the benefit of running code, which may serve as evidence of valuable experimentation and feedback that have made the implemented protocols more mature. It is up to the individual working groups to use this information as they see fit".

6.1. Nokia Implementation

* Organization: Nokia

* Implementation: NSP PCE and SROS PCC.

* Description: Implementation for calculation and conveying intention described in this document

* Maturity Level: Demo

* Coverage: Full

* Contact: andrew.stone@nokia.com

6.2. Cisco Implementation

* Organization: Cisco Systems, Inc.

* Implementation: IOS-XR PCE and PCC.

* Description: Implementation for calculation and conveying intention described in this document

* Maturity Level: Demo

* Coverage: Full

* Contact: ssidor@cisco.com

7. Security Considerations

This document clarifies the behaviour of an existing flag and introduces a new flag to provide further control of that existing behaviour. The introduction of this new flag and behaviour clarification does not create any new sensitive information. No additional security measure is required.

Securing the PCEP session using Transport Layer Security (TLS) [RFC8253], as per the recommendations and best current practices in [RFC9325] is RECOMMENDED.

8. IANA Considerations

[RFC Editor Note: The text below assumes the E bit remains the early allocation value 6. Please adjust if this changes and remove this note before publication.]

This document defines a new bit value in the sub-registry "LSPA Object Flag Field" in the "Path Computation Element Protocol (PCEP) Numbers" registry. IANA has made the following codepoint allocation.

Bit	Name	Reference
6	Protection Enforcement	This document

9. References

9.1. Normative References

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <https://www.rfc-editor.org/info/rfc2119>.

[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>.

[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>.

[RFC3209] Awduche, D., Berger, L., Gan, D., Li, T., Srinivasan, V., and G. Swallow, "RSVP-TE: Extensions to RSVP for LSP Tunnels", RFC 3209, DOI 10.17487/RFC3209, December 2001, <https://www.rfc-editor.org/info/rfc3209>.

[RFC4090] Pan, P., Ed., Swallow, G., Ed., and A. Atlas, Ed., "Fast Reroute Extensions to RSVP-TE for LSP Tunnels", RFC 4090, DOI 10.17487/RFC4090, May 2005, <https://www.rfc-editor.org/info/rfc4090>.

[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>.

[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>.

[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>.

[RFC9325] Sheffer, Y., Saint-Andre, P., and T. Fossati, "Recommendations for Secure Use of Transport Layer Security (TLS) and Datagram Transport Layer Security (DTLS)", BCP 195, RFC 9325, DOI 10.17487/RFC9325, November 2022, <https://www.rfc-editor.org/info/rfc9325>.

9.2. Informative References

[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>.

[RFC7942] Sheffer, Y. and A. Farrel, "Improving Awareness of Running Code: The Implementation Status Section", BCP 205, RFC 7942, DOI 10.17487/RFC7942, July 2016, <https://www.rfc-editor.org/info/rfc7942>.

[RFC8664] Sivabalan, S., Filsfils, C., Tantsura, J., Henderickx, W., and J. Hardwick, "Path Computation Element Communication Protocol (PCEP) Extensions for Segment Routing", RFC 8664, DOI 10.17487/RFC8664, December 2019, <https://www.rfc-editor.org/info/rfc8664>.

[RFC9085] Previdi, S., Talaulikar, K., Ed., Filsfils, C., Gredler, H., and M. Chen, "Border Gateway Protocol - Link State (BGP-LS) Extensions for Segment Routing", RFC 9085, DOI 10.17487/RFC9085, August 2021, <https://www.rfc-editor.org/info/rfc9085>.

Acknowledgements

Thanks to Dhruv Dhody, Mike Koldychev, and John Scudder for reviewing and providing very valuable feedback and discussions on this document.

Thanks to Julien Meuric for shepherding this document.

Authors' Addresses

Andrew Stone
Nokia
600 March Road
Kanata Ontario K2K 2T6
Canada
Email: andrew.stone@nokia.com

Mustapha Aissaoui
Nokia
600 March Road
Kanata Ontario K2K 2T6
Canada
Email: mustapha.aissaoui@nokia.com

Samuel Sidor
Cisco Systems, Inc.
Eurovea Central 3.
Pribinova 10
811 09 Bratislava
Slovakia
Email: ssidor@cisco.com

Siva Sivabalan
Ciena Coroporation
385 Terry Fox Drive
Kanata Ontario K2K 0L1
Canada
Email: ssivabal@ciena.com

pce
Internet-Draft
Intended status: Standards Track
Expires: September 8, 2022

H. Li
A. Wang
China Telecom
Z. Zhang
Juniper Networks
H. Chen
Futurewei
R. Chen
ZTE Corporation
March 07, 2022

Multicast Tree Setup via PCEP
draft-li-pce-multicast-00

Abstract

Multicast forwarding requires per-tree state on certain nodes. Even with BIER, per-tree state is needed on ingress/egress BIER routers (though not needed on transit BIER routers). This documents specifies PCEP protocol to collect tree information (e.g. root, leaf, constraints) to allow a PCE to calculate a tree, and the procedures to set up per-tree forwarding state on relevant nodes for various multicast trees and various replication technologies.

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 September 8, 2022.

Table of Contents

1. Introduction

Currently, multicast management information is mainly signaled by PIM [RFC2362], mLDP [RFC6388] or BGP [RFC6514]. The trees/tunnels are set up using the "receiver-initiated join" technique of PIM/mLDP, hop by hop from downstream routers towards the root. The BGP messages are either sent hop by hop between downstream routers and their upstream neighbors, or can be reflected by Route Reflectors (RRs). The signaling is the interaction between routers and cannot be managed in a unified manner.

As an alternative to each hop independently determining its upstream router and signaling upstream towards the root (following PIM/mLDP model), the entire tree can be calculated by a centralized controller, and the signaling can be entirely done from the controller.

[RFC4655] defines a stateful PCE to be one in which the PCE maintains "strict synchronization between the PCE and not only the network states (in term of topology and resource information), but also the set of computed paths and reserved resources in use in the network." [RFC8231] specifies a set of extensions to PCEP to support state synchronization between PCCs and PCEs.

The controller can serve as a PCE to centrally manage multicast trees and establish states on all tree nodes serving as PCC. This document defines PCEP objects and TLVs to accomplish the multicast source registration/revocation, receiver discovery, multicast tree state update etc. functions.

2. Overview

2.1. Different Multicast Trees

This section discusses various multicast trees like IP, mLDP, SR-P2MP - each with its own tree identification.

2.1.1. IP Multicast

As stated in [RFC1112], an IP multicast packet's destination address is a Multicast Group Address and it follows either a source specific tree rooted at the First Hop Router (FHR) or a shared tree rooted at a Rendezvous Point (RP) [RFC7761]. Each tree is identified by a (source address, group address) pair, where the source address MAY be a wildcard in case of the shared tree. This identification, which is often referred to as (s, g) or (*, g), is used in both forwarding plane as well as in control plane that sets up the tree, e.g., using Protocol Independ Multicast (PIM) [RFC7761] [RFC5015].

Each node on the tree needs to have corresponding forwarding plane state used to forward corresponding traffic and control plane state that is used to set up a tree. Depending on the number of (s, g) and/or (*, g) trees that a routing domain need to support, serious scaling/convergence problems MAY result.

2.1.2. mLDP/RSVP-TE P2MP Tunnels

Quite often, some different multicast trees are used for similar purposes and the tree structure are identical or very similar. For example, in an IPTV application, many channels MAY need to be sent from the same headend router to the same set of edge routers close to receivers.

In that case, a set of IP multicast trees can be transported over a fewer set of MPLS P2MP tunnels - either mLDP [RFC6388] or RSVP-TE P2MP [RFC4875] tunnels - across an MPLS domain. Inside the MPLS domain, there is no IP multicast tree state but only fewer state for the P2MP tunnels. Outside the MPLS domain, those IP multicast trees still have their per-tree state on each tree node.

An mLDP tree is identified by an mLDP FEC in the control plane. In the context of PCEP signaling, part of or an entire mLDP tunnel can be set up using PCEP from a PCE instead of using hop-by-hop mLDP signaling. In this case the only relevance to mLDP is that mLDP FEC is used to identify the tunnel.

A RSVP-TE P2MP tree is identified by a RSVP Session object in the control plane. While in theory it can also be set up via PCEP signaling from a PCE, it is outside the scope of this document.

In the forwarding plane, there is no difference between an mLDP tunnel and a RSVP-TE P2MP tunnel. A tree is identified by a label - incoming labeled traffic is replicated to a bunch of downstream nodes with a label that identifies the tree to the downstream nodes.

2.1.3. SR-P2MP Tunnels

Segment Routing (SR) [RFC8402] has two principles:

1. No per-flow/tunnel state inside an SR domain

2. Optional but perferred use of controllers

When it comes to multicast, the only technologies that sticks to principle #1 is BIER [RFC8279], which does not use per-tree state for forwarding.

SR-P2MP [I-D.ietf-pim-sr-p2mp-policy] [I-D.ietf-spring-sr-replication-segment] sticks to principle #2 only, in that PCEs are used to calculate a multicast tree subject to constraints and then direct signaling from PCEs are used to set up the tree instead of using mLDP/RSVP signaling, but per-tree state is still used.

In the control plane, an SR-P2MP is identified by a (root-id, tree-id) tuple. In an SR-MPLS forwarding plane, an SR-P2MP tunnel is not different from mLDP/RSVP-TE P2MP tunnel at all, though the tree-identifying label is called a tree sid.

In an SRv6 forwarding plane, the tree sid is embedded as part of an IPv6 address.

2.2. Different Replication Technologies

For any kind of multicast trees mentioned above, on a particular tree node A, an incoming packet needs to replicated to a set of downstream nodes B/C/D/E. Note that the upstream and downstream nodes MAY be directly connected or they can be reached via a set of P2P/MP2P tunnels, P2MP tunnels, or via BIER. In the latter case, intermediate nodes do not maintain the per-tree state but they do need to maintain state for the transporting tunnels or BIER.

Now on node A/B/C/D/E, per-tree replication state needs to be set up. In particular on A, depending on the method used to replicate to

B/C/D/E, (a combination of) the following replication branches are needed:

- o A set of outgoing interfaces for native IP forwarding to directly connected nodes from the B/C/D/E set in case of IP multicast tree, and/or,

- o A set of P2P/MP2P/P2MP tunnels to reach indirectly connected nodes from the B/C/D/E set, and/or,

- o A BIER bitstring and relevant BIER information to reach some nodes from the B/C/D/E set

In the latter two cases, the replication branches also need relevant information to identify to B/C/D/E which tree a packet is for (e.g., a label for an mLDP or SR-P2MP tree).

2.3. Tree Information Discovery

When a PCE calculates a multicast tree, it needs to collect the tree information like root, leaves and calculation constraints. This MAY be provided to the PCE via a southbound (wrt the PCE) interface from an orchestrator or via northbound PCEP signaling from some PCC nodes (e.g. the root and leaf nodes of the tree). The PCE MAY also participate overlay signaling protocols to collect the information (e.g., [I-D.zzhang-mvpn-evpn-controller]).

This document only covers the northbound PCEP signaling. The two other methods mentioned above are out of scope of this document.

Whether it is an IP Multicast or mLDP/SR-P2MP Tree, the root, leaves and calculation constraints are encoded the same way but associated with different tree identifications in the PCEP signaling. Specifically, the root and leaves are encoded as IP addresses.

The leaf information can be encoded as a full set of leaves or as addition/removal delta.

2.3.1. Root node Discovery

When a multicast source accesses to a First Hop Router (FHR), the FHR originates a PCRpt message carrying the Multicast Source Registration (MSR) object defined in Section 3.3, which provides FHR information. FHR, as PCC, delegate the controller as PCE to calculate multicast tree. In the sent PCRpt message, the D bit of LSP Object is set to 1.

2.3.2. Leaf node Discovery

For IP multicast, Last Hop Router (LHR) MAY convert the multicast source and group address carried in the IGMP Join/Leave messages sent by local hosts and VPN information corresponding to local hosts into PCRpt messages and report these information to the controller. Similar to IGMP messages, PIM messages from other domains can also be used as receiver information and be converted into PCRpt messages to be reported to the controller.

For mLDP and SR-P2MP multicast, The corresponding tree identifiers can also be carried in PCRpt messages and reported to the controller.

When the first receiver accesses a Last Hop Router (LHR) to join a multicast tree, or when the last receiver leaves the LHR, the LHR originates a PCRpt message carrying the Multicast Receiver Information (MRI) object defined in Section 3.4, which provides LHR information.

If the receivers locate in different VPN domains, the VPN information associated with the multicast source and multicast destination should also be reported by the LHR and FHR. The VPN information reported by LHRs SHOULD be consistent.

2.4. Multicast Tree

The multicast trees are established with the FHRs of the multicast source as the root of the multicast trees. According to different multicast technologies, the types of multicast trees include mLDP tree, SR-P2MP tree and BIER tree. Different multicast trees correspond to different processing. This section describes the state update of different multicast operations.

No matter which forwarding technology is used, there is a case that local multicast receivers directly access to FHRs. In this case, FHRs need to forward multicast data for local receivers in the way of native IP without other control information.

2.4.1. mLDP Tree

For a multicast tree, each node needs a tree label to identify a multicast tree. There are three options for tree label assignment:

- From each router’s SRLB that the controller learns
- From the common SRGB that the controller learns
- From the controller’s local label space

The detailed instruction is as per [I-D.ietf-bess-bgp-multicast-controller]. Section 3.5.1 defines a new TLV to extend the CCI Object-Type support the tree label.

The operations for initiating multicast tree LSPs and downloadinglabels are consistent with [I-D.ietf-pce-sr-p2mp-policy].

2.4.2. SR-P2MP Tree

The processing of PCE-based SR P2MP policy delivery and P2MP path establishment has been defined in [I-D.ietf-pce-sr-p2mp-policy]. For the calculated multicast tree, SR P2MP multicast tree update can refer to that document.

2.4.3. BIER Tree

Different from other forwarding technologies, BIER does not need the transit nodes to maintain the multicast state, and only needs to forward and encapsulate the packets according to the BitString and BIFT. BIFT information can be learned by IGP. For each node of a sub-domain, the BIFT can be configured locally or allocated by BGP controller or PCE. Allocation by PCE is as per [I-D.chen-pce-pcep-extension-pce-controller-bier].

After receiving the BFR information reported from LHRs, the controller combine BFR information of LHRs into a BitString to guide multicast data forwarding.

The controller (PCE) sends the BitString to the FHR via PCInitiate or PCUpd message carrying Tree Forwarding State Synchronization (TFSS) object defined in Section 3.6 to inform the FHR to forward multicast data for a specific multicast tree according to the BitString, and the tree identifier is (*, g) or (s, g) tuple.

2.5. Multicast Statistics

Multicast statistics are helpful for multicast service analysis and business development. This section describes how to collect statistics about multicast users.

For each LHR, the statistics consist of two parts, one is the number of local users in the its managed domain, and the other is the number of other managed domains. The sum of these two parts is the number of multicast data that the one LHR needs to duplicate.

LHRs send this information to the controller via PCRpt message carrying Multicast Receiver Statistics (MRS) object. LHRs need report this information to the controller periodically. Once the

statistics information of a LHR change, the LHR also needs reports to the controller.

Controller collects the statistics reported by LHRs and MAY periodically informs the FHR of the statistics via PCRpt messages carrying MRS object so that the FHR can feed the statistics back to the multicast source.

3. PCEP Object Formats

3.1. OPEN Object

3.1.1. STATEFUL-PCE-CAPABILITY TLV

During the PCEP initialization phase, PCEP speakers advertise stateful capability via the STATEFUL-PCE-CAPABILITY TLV in the OPEN object. Various flags are defined for the STATEFUL-PCE-CAPABILITY TLV defined in [RFC8231] and updated in [RFC8232] and [RFC8281].

A new flag is added in this document, whose code point is TBD1:

MULTICAST-STATE-CAPABILITY bit, if this bit is set to 1 by a PCEP speaker, it indicates that the PCEP speaker supports the capability of these new extensions as specified in this document.

To support the multicast tree state management capabilities described in this document, both PCE and PCC need to set MULTICAST-STATE-CAPABILITY bit. If a PCEP speaker receivers PECP message with the newly defined object, but without the MULTICAST-STATE-CAPABILITY bit set in STATEFUL-PCE-CAPABILITY TLV in the OPEN object, it MUST:

- o Send a PCErr message with Error-Type=10(Reception of an invalid object) and Error-Value TBD2(MULTICAST-STATE-CAPABILITY bit is not set).

- o Terminate the PCEP session.

3.1.2. PCECC Capability sub-TLV

The PCECC-CAPABILITY sub-TLV is an optional TLV for use in the OPEN Object for PCECC capability advertisement in PATH-SETUP-TYPE-CAPABILITY TLV as specified in [RFC9050].

[I-D.dhody-pce-pcep-extension-pce-controller-p2mp] adds a new flag (M Bit) in the PCECC-CAPABILITY sub-TLV to indicate the support for P2MP in PCECC, which is applicable for multicast tree described in this document as well.

3.2. PATH-SETUP-TYPE TLV

The PATH-SETUP-TYPE TLV is defined in [RFC8408]; [RFC9050] defines a PST value for PCECC as '2', which is applicable for multicast tree described in this document as well.

3.3. Multicast Source Registration Object

The MSR object is optional and SHOULD contain multicast source information and FHR information. The MSR object SHOULD be carried within a PCRpt message sent by PCC to PCE for registration. The MSR object SHOULD be carried within a PCUpd message sent by PCE to PCC in response to registration.

MSR Object-Class is TBD3. MSR Object-Type is 1. The format of the MSR object body is:

```
 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
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                            flags                            |R|A|
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|        Auxiliary Length       |            Reserved           |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
~                        Auxiliary Data                         ~
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
~                        Optional TLVs                          ~
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 1: MSR Object Body Format

R (Register flag, 1 bit): The R flag set to 1 indicates that the PCC is registering multicast information to the PCE. The R flag set to 0 indicates that the PCC revokes the registration.

A (Authentication flag, 1 bit): The A flag set to 1 indicates success of registration. The A flag set to 0 indicates failure of registration or revocation of registration. If there is no authentication information, A flag SHOULD also be set to 0.

Auxiliary Length (1 Octet): indicates the length of Auxiliary Data.

Auxiliary Data (Variable length): contains functional data such as authentication information.

MSR object MAY include Multicast Address TLV, mLDP FEC TLV, SR-IPV4-P2MP-POLICY-ID TLV, SR-IPV6-P2MP-POLICY-ID TLV, and VPN Information TLV. SR-IPV4-P2MP-POLICY-ID TLV and SR-IPV6-P2MP-POLICY-ID TLV are

defined in [I-D.ietf-pce-sr-p2mp-policy]. MSR object carries different TLVs depending on the multicast tree.

3.3.1. Multicast Address TLV

In IP multicast scenarios, Multicast Address TLV SHOULD be included.

The format of the Multicast Address TLV is:

```
 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 = TBD4          |             Length            |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|         Prefix Length         |            Reserved           |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
~                     Multicast Source Address                  ~
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
~                     Multicast Group Address                   ~
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 2: Multicast Address TLV Format

Prefix Length (2 Octets): indicates the length of multicast addresses. the length MAY be 4 Octets, 8 Octets, 16 Octets, or 32 Octets. The multicast source address can be empty, but the multicast group address must be filled. If both the multicast source address and multicast group address exist, they must be both IPv4 and IPv6 addresses.

Multicast Source Address (Variable length): contains IPv4 or IPv6 address of the multicast source.

Multicast Group Address (Variable length): contains IPv4 or IPv6 address of the multicast group.

3.3.2. mLDP FEC TLV

In mLDP multicast scenarios, Multicast Address TLV SHOULD be included.

The format of the mLDP FEC TLV is:

```
 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          |             Length            |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
~                            mLDP FEC                           ~
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 3: mLDP FEC TLV Format

mLDP FEC (Variable length): carries mLDP FEC.

3.3.3. VPN Information TLV

VPN Information TLV is used to report VPN information about multicast sources and receivers. The format of the VPN Information TLV is:

```
 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          |             Length            |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                         VPN Identifier                        |
|                                                               |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 4: VPN Information TLV Format

VPN Identifier (8 Octets): indicates the VPN which the receiver used.

3.4. Multicast Receiver Information Object

The MRI object is optional and SHOULD contain receivers' information for matching the multicast registration information. The MRI object SHOULD be carried within a PCRpt message sent by PCC to PCE for multicast joining or leaving.

MRI Object-Class is TBD7. MRI Object-Type is 1. The format of the MRI object body is:

```
 0                   1                   2                   3
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|            Reserved           |            flags          |B|S|
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
~                        Optional TLVs                          ~
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 5: MRI Object Body Format

B(BIER multicast flag, 1 bit): The B flag set to 1 indicates that multicast protocol is BIER. If the B flag set to 1, MRI object SHOULD include BFR information TLV defined in Section 3.4.1.

S(Subscribe flag, 1 bit): The S flag set to 1 indicates that PCC delivers the message requesting to join PCE. The S flag set to 0 indicates that PCC delivers the message requesting to leave to PCE.

MRI object MAY include Multicast Address TLV, mLDP FEC TLV, SR-IPV4-P2MP-POLICY-ID TLV, SR-IPV6-P2MP-POLICY-ID TLV, VPN Information TLV, and BFR information TLV.

3.4.1. BFR Information TLV

BFR Information TLV is used to report router location information in the BIER domain. The format of the BFR Information TLV is:

```
 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 = TBD8          |             Length            |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
| Subdomain-Id  |            BFR-ID             |  BSL  |Padding|
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 6: BFR Information TLV Format

Subdomain-Id (1 Octet): Unique value identifying the BIER subdomain.

BFR-ID (2 Octets): Identification of BFR in a subdomain.

BSL (BitString Length, 4 bits): encodes the length in bits of the BitString as per [RFC8296], the maximum length of the BitString is 7, it indicates the length of BitString is 4096. It is used to refer to the number of bits in the BitString.

3.5. CCI Object

The CCI object [RFC9050] is used by the PCE to specify the forwarding instructions to the PCC and MAY be carried within a PCInitiate or PCRpt message for control information download/report.

The CCI Object Type 1 for MPLS Label is defined in [RFC9050], which is used for the P2MP LSPs as well. For mLDP multicast trees, in addition to forwarding labels, corresponding tree labels or label stacks are required. Therefore, a new TLV is needed to carry tree labels or label stacks.

3.5.1. MPLS Tree Label TLV

The format of MPLS Tree Label TLV is:

```
 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 = TBD9          |             Length            |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
~                     MPLS Tree Label stack                     ~
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 7: MPLS Tree Label TLV Format

MPLS Tree Label stack (Variable length): contains tree labels used to identify multicast trees. There are three application scenarios, as per [I-D.ietf-bess-bgp-multicast-controller].

3.6. Tree Forwarding State Synchronization Object

The TFSS object is optional and SHOULD contain the forwarding state of control plane that the tree node needs to synchronize. The TFSS object SHOULD be carried within a PCInitiate or PCUpd message sent by PCE to PCC, or a PCRpt message sent by PCC to PCE for response.

TFSS Object-Class is TBD10. TFSS Object-Type is 1. The format of the TFSS object body is:

```
 0                   1                   2                   3
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|            Reserved           |              flags          |F|
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
~                        Optional TLVs                          ~
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 8: TFSS Object Body Format

F (Forwarding flag, 1 bit): The F flag set to 1 indicates that the node MAY start forwarding multicast packets. The F flag set to 0 indicates that the node SHOULD stop forwarding multicast packets.

TFSS object MAY include Multicast Address TLV, VPN Information TLV, and BIER Attribute TLV. When TT=1, BIER Attribute TLV SHOULD be included.

3.6.1. BIER Attribute TLV

BIER Attribute TLV is used to instruct the FHR to forward multicast packets to the users according to the BitString specified by the controller. The format of BIER Attribute TLV is:

```
 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 = TBD11         |             Length            |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
| Subdomain-id  |       SI      |  BSL  ~       BitString       ~
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 9: BIER Attribute TLV Format

Subdomain-id (1 Octet): Unique value identifying the BIER subdomain.

SI (Set Identifier, 1 Octet): encoding the Set Identifier used in the encapsulation for this BIER subdomain for this BitString length.

BSL (BitString Length, 4 bits): encodes the length in bits of the BitString as per [RFC8296], the maximum length of the BitString is 7, it indicates the length of BitString is 4096. It is used to refer to the number of bits in the BitString.

BitString(Variable length): indicates the path of multicast data packets forwarding for headend.

3.7. Multicast Receiver Statistics Object

The MRS object is optional and used to inform PCE of the number of receivers. The MRS object SHOULD be carried within a PCRpt or a PCUpd message for synchronize receiver information periodically, or PCRpt message for the leaving of receivers.

MRS Object-Class is TBD12. MRS Object-Type is 1. The format of the MRS object body is:

```
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
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|          Number Length        |            Reserved           |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                      Multicast Statistics                     |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
~                        Optional TLVs                          ~
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 10: MRS Object Body Format

Number Length (2 Octets): indicates the length of receiver number.

Multicast Statistics (4 Octets): indicates the statistics information for a particular multicast tree of a controller or a LHR.

MRS object MAY include Multicast Address TLV, mLDP FEC TLV, SR-IPV4-P2MP-POLICY-ID TLV, and SR-IPV6-P2MP-POLICY-ID TLV.

4. Message Formats

4.1. PCRpt message

The definition of the PCRpt message from [RFC8231] and [RFC9050] is extended to optionally include MSR object, MRI object, TFSS object, and MRS object after the path object. The encoding will become:

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

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

<state-report> ::= [<SRP>]
                   <LSP>
                   <path>
                   [<MSR>]
                   [<MRI>]
                   [<TFSS>]
                   [<MRS>]


Where:
  <path> is as per [RFC8231] and the LSP and SRP object are
  also defined in [RFC8231].
```

4.2. PCInitiate message

The definition of the PCInitiate message from [RFC8281] is extended to optionally include TFSS object after the path object. The encoding will become:

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

Where:

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

<update-request> ::= <SRP>
                     <LSP>
                     [<TFSS>]
```

Where:
LSP and SRP object are defined in [RFC8231].

4.3. PCUpd message

The definition of the PCUpd message from [RFC8231] is extended to optionally include MSR object, TFSS object and MRS object after the path object. The encoding will become:

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

Where:

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

<update-request> ::= <SRP>
                     <LSP>
                     <path>
                     [<MSR>]
                     [<TFSS>]
                     [<MRS>]
```

Where:
<path> is as per [RFC8231] and the LSP and SRP object are also defined in [RFC8231].

5. Example Workflow

5.1. Multicast Tree Discovery

```
                 +-------+                       +-------+
                 |PCC    |                       |  PCE  |
                 |ingress|                       +-------+
           +-----|       |                           |
           |PCC  +-------+                           |
           |transit| |                               |
    +------|       | |                               |
    |PCC   +-------+ |                               |
    |egress |  |     |                               |
    +-------+  |     |                               |
        |      |     |-----PCRpt,PLSP-ID=1,D=1------>|Source Registration
        |      |     |       MSR Object(R=1)         |(Root Discovery)
        |      |     |                               |
        |      |     |<----PCUpd,PLSP-ID=1-----------|Source Registration
        |      |     |     MSR Object(R=1)           |Recognition
        |      |     |                               |
        |------------------PCRpt,PLSP-ID=0---- ----->|Receiver Report
        |      |     |     MRI Objetc(S=1)           |(Leaf Discovery)
        |      |     |                               |
```

Figure 11: Workflow of Multicast Tree Discovery

5.2. Multicast Tree Path Establishment and Update

5.2.1. mLDP Tree and SR-P2MP Tree

The workflow of mLDP Tree and SR-P2MP Tree are as per [I-D.ietf-pce-sr-p2mp-policy].

5.2.2. BIER Tree

```
                  +-------+                        +-------+
                  |PCC    |                        |  PCE  |
                  |ingress|                        +-------+
          +------|       |                            |
          |      +-------+                            |
          | transit| |                               |
    +------|        | |                               |
    |      +--------+ |                               |
    |egress  |  |     |                               |
    +--------+  |     |                               |
        |       |     |<-----PCInitiate,PLSP-ID=1-------|Tree State
        |       |     |      TFSS Objetc (F=1,TT=1)     |Setup
        |       |     |                               |
        |       |     |------PCRpt,PLSP-ID=1----------->|Tree State
        |       |     |                               |Recognition
        |       |     |                               |
        |--------------------PCRpt,PLSP-ID=0----------->|Receiver Report
        |       |     |      MRI Objetc(S=1/S=0)        |(Leaf Discovery)
        |       |     |                               |
        |       |     |<-----PCUpd,PLSP-ID=1------------|Tree State
        |       |     |      TFSS Objetc (F=1,TT=1)     |Update
        |       |     |                               |
        |       |     |------PCRpt,PLSP-ID=1----------->|Tree State
        |       |     |                               |Recognition
        |       |     |                               |
```

Figure 12: Workflow of Multicast Tree Discovery

In the previous multicast tree discovery process, ingress has delegated the PCE to calculate the multicast path. Therefore, during the establishment of the multicast tree, PCE informs the ingress to forward multicast data for (S,G)/(*,G) via PCInitiate message carrying TFSS object according to the delegation. The PLSP-ID is the value specified by the ingress.

Ingress then updates the tree state and responds to the PCE. At this point, the BIER tree is formally established and the ingress starts forwarding for the multicast according to the BitString specified by PCE.

If the topology of the BIER tree changes, the corresponding egresses will send PCRpt messages carring MRI object to PCE to inform PCE of their changes.

The PCE updates the BitStrig based on changes of the egresses and informs the ingress to update. State update process is similar to state setup process, except for the type of message sent by the PCE.

6. Security Considerations

To be added.

7. IANA Considerations

7.1. MULTICAST-STATE-CAPABILITY

IANA is requested to allocate a new code point within registry "STATEFUL-PCE-CAPABILITY TLV Flag Field" under "Path Computation Element Protocol (PCEP) Numbers" as follows:

Value	Description	Reference
TBD1	MULTICAST-STATE-CAPABILITY	This document

7.2. PCEP-ERROR Object

IANA is requested to allocate code-points in the "PCEP-ERROR Object Error Types and Values" sub-registry for the following new error-type and error-value:

Error-Type	Description	Reference
10	Error-value = TBD2 MULTICAST-STATE-CAPABILITY bit is not set	This document

7.3. New Objects

IANA is requested to allocate the following Object-Class Values in the "PCEP Objects" sub-registry under the "Path Computation Element Protocol (PCEP) Numbers" registry:

Object-Class Value	Description	Reference
TBD3	Multicast Source Registration	This document
TBD7	Multicast Receiver Information	This document
TBD10	Tree Forwarding State Synchronization	This document
TBD12	Multicast Receiver Statistics	This document

IANA is requested to allocate the following Object-Type Value of CCI object in the "PCEP Objects" sub-registry under the "Path Computation Element Protocol (PCEP) Numbers" registry:

7.4. New TLVs

IANA is requested to allocate the following TLVs:

Type	Description	Reference
TBD4	Multicast Source Address	This document
TBD5	mLDP FEC	This document
TBD6	VPN Information	This document
TBD8	BFR Information	This document
TBD9	MPLS Tree Label	This document
TBD10	BIER Attribute	This document

8. Acknowledgements

The author thanks ... for their review and comments.

9. References

9.1. Normative References

[I-D.ietf-pce-sr-p2mp-policy]
Bidgoli, H., Voyer, D., Rajarathinam, S., Hemmati, E., Saad, T., and S. Sivabalan, "PCEP extensions for p2mp sr policy", draft-ietf-pce-sr-p2mp-policy-00 (work in progress), December 2021.

[RFC1112] Deering, S., "Host extensions for IP multicasting", STD 5, RFC 1112, DOI 10.17487/RFC1112, August 1989, <https://www.rfc-editor.org/info/rfc1112>.

[RFC2362] Estrin, D., Farinacci, D., Helmy, A., Thaler, D., Deering, S., Handley, M., Jacobson, V., Liu, C., Sharma, P., and L. Wei, "Protocol Independent Multicast-Sparse Mode (PIM-SM): Protocol Specification", RFC 2362, DOI 10.17487/RFC2362, June 1998, <https://www.rfc-editor.org/info/rfc2362>.

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

[RFC4875] Aggarwal, R., Ed., Papadimitriou, D., Ed., and S. Yasukawa, Ed., "Extensions to Resource Reservation Protocol - Traffic Engineering (RSVP-TE) for Point-to-Multipoint TE Label Switched Paths (LSPs)", RFC 4875, DOI 10.17487/RFC4875, May 2007, <https://www.rfc-editor.org/info/rfc4875>.

[RFC5015] Handley, M., Kouvelas, I., Speakman, T., and L. Vicisano, "Bidirectional Protocol Independent Multicast (BIDIR-PIM)", RFC 5015, DOI 10.17487/RFC5015, October 2007, <https://www.rfc-editor.org/info/rfc5015>.

[RFC6388] Wijnands, IJ., Ed., Minei, I., Ed., Kompella, K., and B. Thomas, "Label Distribution Protocol Extensions for Point-to-Multipoint and Multipoint-to-Multipoint Label Switched Paths", RFC 6388, DOI 10.17487/RFC6388, November 2011, <https://www.rfc-editor.org/info/rfc6388>.

[RFC6514] Aggarwal, R., Rosen, E., Morin, T., and Y. Rekhter, "BGP Encodings and Procedures for Multicast in MPLS/BGP IP VPNs", RFC 6514, DOI 10.17487/RFC6514, February 2012, <https://www.rfc-editor.org/info/rfc6514>.

[RFC7761] Fenner, B., Handley, M., Holbrook, H., Kouvelas, I., Parekh, R., Zhang, Z., and L. Zheng, "Protocol Independent Multicast - Sparse Mode (PIM-SM): Protocol Specification (Revised)", STD 83, RFC 7761, DOI 10.17487/RFC7761, March 2016, <https://www.rfc-editor.org/info/rfc7761>.

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

[RFC8232] Crabbe, E., Minei, I., Medved, J., Varga, R., Zhang, X., and D. Dhody, "Optimizations of Label Switched Path State Synchronization Procedures for a Stateful PCE", RFC 8232, DOI 10.17487/RFC8232, September 2017, <https://www.rfc-editor.org/info/rfc8232>.

[RFC8279] Wijnands, IJ., Ed., Rosen, E., Ed., Dolganow, A., Przygienda, T., and S. Aldrin, "Multicast Using Bit Index Explicit Replication (BIER)", RFC 8279, DOI 10.17487/RFC8279, November 2017, <https://www.rfc-editor.org/info/rfc8279>.

[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>.

[RFC8296] Wijnands, IJ., Ed., Rosen, E., Ed., Dolganow, A., Tantsura, J., Aldrin, S., and I. Meilik, "Encapsulation for Bit Index Explicit Replication (BIER) in MPLS and Non-MPLS Networks", RFC 8296, DOI 10.17487/RFC8296, January 2018, <https://www.rfc-editor.org/info/rfc8296>.

[RFC8402] Filsfils, C., Ed., Previdi, S., Ed., Ginsberg, L., Decraene, B., Litkowski, S., and R. Shakir, "Segment Routing Architecture", RFC 8402, DOI 10.17487/RFC8402, July 2018, <https://www.rfc-editor.org/info/rfc8402>.

[RFC9050] Li, Z., Peng, S., Negi, M., Zhao, Q., and C. Zhou, "Path Computation Element Communication Protocol (PCEP) Procedures and Extensions for Using the PCE as a Central Controller (PCECC) of LSPs", RFC 9050, DOI 10.17487/RFC9050, July 2021, <https://www.rfc-editor.org/info/rfc9050>.

9.2. Informative References

[I-D.chen-pce-pcep-extension-pce-controller-bier]
Chen, R. and B. Xu, "PCEP Procedures and Protocol Extensions for Using PCE as a Central Controller (PCECC) of BIER", draft-chen-pce-pcep-extension-pce-controller-bier-02 (work in progress), December 2021.

[I-D.dhody-pce-pcep-extension-pce-controller-p2mp]
Li, Z., Peng, S., Geng, X., and M. S. Negi, "Path Computation Element Communication Protocol (PCEP) Procedures and Extensions for Using the PCE as a Central Controller (PCECC) of point-to-multipoint (P2MP) LSPs", draft-dhody-pce-pcep-extension-pce-controller-p2mp-08 (work in progress), March 2022.

[I-D.ietf-bess-bgp-multicast-controller]
Zhang, Z., Raszuk, R., Pacella, D., and A. Gulko, "Controller Based BGP Multicast Signaling", draft-ietf-bess-bgp-multicast-controller-07 (work in progress), July 2021.

[I-D.ietf-pim-sr-p2mp-policy]
(editor), D. V., Filsfils, C., Parekh, R., Bidgoli, H., and Z. Zhang, "Segment Routing Point-to-Multipoint Policy", draft-ietf-pim-sr-p2mp-policy-03 (work in progress), August 2021.

[I-D.ietf-spring-sr-replication-segment]
(editor), D. V., Filsfils, C., Parekh, R., Bidgoli, H., and Z. Zhang, "SR Replication Segment for Multi-point Service Delivery", draft-ietf-spring-sr-replication-segment-06 (work in progress), October 2021.

[I-D.zzhang-mvpn-evpn-controller]
Zhang, Z., Parekh, R., Zhang, Z., and H. Bidgoli, "MVPN and EVPN BUM Signaling with Controllers", draft-zzhang-mvpn-evpn-controller-01 (work in progress), October 2021.

Authors' Addresses

Huanan Li
China Telecom

Email: lihn6@foxmail.com

Aijun Wang
China Telecom

Email: wangaj3@chinatelecom.cn

Zhaohui Zhang
Juniper Networks

Email: zzhang@juniper.net

Huaimo Chen
Futurewei

Email: Huaimo.chen@futurewei.com

Ran Chen
ZTE Corporation

Email: chen.ran@zte.com.cn

pce
Internet-Draft
Intended status: Standards Track
Expires: 24 March 2023

H. Li
A. Wang
China Telecom
Z. Zhang
Juniper Networks
H. Chen
Futurewei
R. Chen
ZTE Corporation
20 September 2022

Multicast Tree Setup via PCEP
draft-li-pce-multicast-02

Abstract

Multicast forwarding requires per-tree state on certain nodes. Even with BIER, per-tree state is needed on ingress/egress BIER routers (though not needed on transit BIER routers). This documents specifies PCEP protocol to collect tree information (e.g. root, leaf, constraints) to allow a PCE to calculate a tree, and the procedures to set up per-tree forwarding state on relevant nodes for various multicast trees and various replication technologies.

Status of This Memo

Copyright Notice

Table of Contents

1. Introduction

Currently, during the establishment of multicast services, multicast management information is mainly signaled by PIM [RFC2362], mLDP [RFC6388] or BGP [RFC6514]. The trees/tunnels are set up using the "receiver-initiated join" technique of PIM/mLDP, hop by hop from downstream routers towards the root. The BGP messages are either sent hop by hop between downstream routers and their upstream neighbors, or can be reflected by Route Reflectors (RRs). The control signaling interaction of multicast information is between routers and cannot be managed in a unified manner.

As an alternative to each hop independently determining its upstream router and signaling upstream towards the root (following PIM/mLDP model), the entire tree can be calculated by a centralized controller, and the signaling can be entirely done from the controller.

[RFC4655] defines a stateful PCE to be one in which the PCE maintains "strict synchronization between the PCE and not only the network states (in term of topology and resource information), but also the set of computed paths and reserved resources in use in the network." [RFC8231] specifies a set of extensions to PCEP to support state synchronization between PCCs and PCEs.

The controller can serve as a PCE to centrally manage multicast trees and establish states on all tree nodes serving as PCC. This document defines PCEP objects and TLVs to accomplish the multicast source registration/revocation, receiver discovery, multicast tree state update etc. functions.

2. Overview

2.1. Different Multicast Trees

This section discusses various multicast trees like IP, mLDP, SR-P2MP - each with its own tree identification.

2.1.1. IP Multicast

As stated in [RFC1112], an IP multicast packet's destination address is a Multicast Group Address and it follows either a source specific tree rooted at the First Hop Router (FHR) or a shared tree rooted at a Rendezvous Point (RP) [RFC7761]. Each tree is identified by a (source address, group address) pair, where the source address MAY be a wildcard in case of the shared tree. This identification, which is often referred to as (s, g) or (*, g), is used in both forwarding plane as well as in control plane that sets up the tree, e.g., using Protocol Independ Multicast (PIM) [RFC7761] [RFC5015].

Each node on the tree needs to have corresponding forwarding plane state used to forward corresponding traffic and control plane state that is used to set up a tree. Depending on the number of (s, g) and/or (*, g) trees that a routing domain need to support, serious scaling/convergence problems MAY result.

2.1.2. mLDP/RSVP-TE P2MP Tunnels

Quite often, some different multicast trees are used for similar purposes and the tree structure are identical or very similar. For example, in an IPTV application, many channels MAY need to be sent from the same headend router to the same set of edge routers close to receivers.

In that case, a set of IP multicast trees can be transported over a fewer set of MPLS P2MP tunnels - either mLDP [RFC6388] or RSVP-TE P2MP [RFC4875] tunnels - across an MPLS domain. Inside the MPLS domain, there is no IP multicast tree state but only fewer state for the P2MP tunnels. Outside the MPLS domain, those IP multicast trees still have their per-tree state on each tree node.

An mLDP tree is identified by an mLDP FEC in the control plane. In the context of PCEP signaling, part of or an entire mLDP tunnel can be set up using PCEP from a PCE instead of using hop-by-hop mLDP signaling. In this case the only relevance to mLDP is that mLDP FEC is used to identify the tunnel.

A RSVP-TE P2MP tree is identified by a RSVP Session object in the control plane. While in theory it can also be set up via PCEP signaling from a PCE, it is outside the scope of this document.

In the forwarding plane, there is no difference between an mLDP tunnel and a RSVP-TE P2MP tunnel. A tree is identified by a label - incoming labeled traffic is replicated to a bunch of downstream nodes with a label that identifies the tree to the downstream nodes.

2.1.3. SR-P2MP Tunnels

Segment Routing (SR) [RFC8402] has two principles:

1. No per-flow/tunnel state inside an SR domain

2. Optional but perferred use of controllers

When it comes to multicast, the only technologies that sticks to principle #1 is BIER [RFC8279], which does not use per-tree state for forwarding.

SR-P2MP [I-D.ietf-pim-sr-p2mp-policy] [I-D.ietf-spring-sr-replication-segment] sticks to principle #2 only, in that PCEs are used to calculate a multicast tree subject to constraints and then direct signaling from PCEs are used to set up the tree instead of using mLDP/RSVP signaling, but per-tree state is still used.

In the control plane, an SR-P2MP is identified by a (root-id, tree-id) tuple. In an SR-MPLS forwarding plane, an SR-P2MP tunnel is not different from mLDP/RSVP-TE P2MP tunnel at all, though the tree-identifying label is called a tree sid.

In an SRv6 forwarding plane, the tree sid is embedded as part of an IPv6 address.

2.2. Different Replication Technologies

For any kind of multicast trees mentioned above, on a particular tree node A, an incoming packet needs to replicated to a set of downstream nodes B/C/D/E. Note that the upstream and downstream nodes MAY be directly connected or they can be reached via a set of P2P/MP2P tunnels, P2MP tunnels, or via BIER. In the latter case, intermediate nodes do not maintain the per-tree state but they do need to maintain state for the transporting tunnels or BIER.

Now on node A/B/C/D/E, per-tree replication state needs to be set up. In particular on A, depending on the method used to replicate to B/C/D/E, (a combination of) the following replication branches are needed:

* A set of outgoing interfaces for native IP forwarding to directly connected nodes from the B/C/D/E set in case of IP multicast tree, and/or,

* A set of P2P/MP2P/P2MP tunnels to reach indirectly connected nodes from the B/C/D/E set, and/or,

* A BIER bitstring and relevant BIER information to reach some nodes from the B/C/D/E set

In the latter two cases, the replication branches also need relevant information to identify to B/C/D/E which tree a packet is for (e.g., a label for an mLDP or SR-P2MP tree).

2.3. Tree Information Discovery

When a PCE calculates a multicast tree, it needs to collect the tree information like root, leaves and calculation constraints. This MAY be provided to the PCE via a southbound (wrt the PCE) interface from an orchestrator or via northbound PCEP signaling from some PCC nodes (e.g. the root and leaf nodes of the tree). The PCE MAY also participate overlay signaling protocols to collect the information (e.g., [I-D.zzhang-mvpn-evpn-controller]).

This document only covers the northbound PCEP signaling. The two other methods mentioned above are out of scope of this document.

Whether it is an IP Multicast or mLDP/SR-P2MP Tree, the root, leaves and calculation constraints are encoded the same way but associated with different tree identifications in the PCEP signaling. Specifically, the root and leaves are encoded as IP addresses.

The leaf information can be encoded as a full set of leaves or as addition/removal delta.

2.3.1. Root node Discovery

When a multicast source accesses to a First Hop Router (FHR), the FHR originates a PCRpt message carrying the Multicast Source Registration (MSR) object defined in Section 3.3, which provides FHR information. FHR, as PCC, delegate the controller as PCE to calculate multicast tree. In the sent PCRpt message, the D bit of LSP Object is set to 1.

2.3.2. Leaf node Discovery

For IP multicast, Last Hop Router (LHR) MAY convert the multicast source and group address carried in the IGMP Join/Leave messages sent by local hosts and VPN information corresponding to local hosts into PCRpt messages and report these information to the controller. Similar to IGMP messages, PIM messages from other domains can also be used as receiver information and be converted into PCRpt messages to be reported to the controller.

For mLDP and SR-P2MP multicast, The corresponding tree identifiers can also be carried in PCRpt messages and reported to the controller.

When the first receiver accesses a Last Hop Router (LHR) to join a multicast tree, or when the last receiver leaves the LHR, the LHR originates a PCRpt message carrying the Multicast Receiver Information (MRI) object defined in Section 3.4, which provides LHR information.

If the receivers locate in different VPN domains, the VPN information associated with the multicast source and multicast destination should also be reported by the LHR and FHR. The VPN information reported by LHRs SHOULD be consistent.

2.4. Multicast Tree

The multicast trees are established with the FHRs of the multicast source as the root of the multicast trees. According to different multicast technologies, the types of multicast trees include labeled tree and BIER tree. Different multicast trees correspond to different processing. This section describes the state update of different multicast operations.

No matter which forwarding technology is used, there is a case that local multicast receivers directly access to FHRs. In this case, FHRs need to forward multicast data for local receivers in the way of native IP without other control information.

In some scenarios, the egresses node needs a VPN forwarding identifier to determine which VRF to forward the packet to. The allocation of VPN forwarding identifier is performed by controller. The scenario of egress assigning VPN tags has not been considered yet. Controller sends the VPN forwarding identifier to ingress and instructs ingress to encapsulate it. The VPN Forwarding Label may be MPLS label or SRv6 segment.

2.4.1. Labeled Tree

For a Labeled multicast tree, whether the forwarding is based on native IP, mLDP or SR-P2MP tunnel, each node needs a tree label to identify a multicast tree. There are three options for tree label assignment:

* From each router's SRLB that the controller learns

* From the common SRGB that the controller learns

* From the controller's local label space

The detailed instruction is as per [I-D.ietf-bess-bgp-multicast-controller]. Section 3.5.1 defines two new TLVs to extend the CCI Object-Type support the tree label and VPN Label.

The operations for initiating multicast tree LSPs and downloadinglabels are consistent with [I-D.ietf-pce-sr-p2mp-policy].

2.4.2. BIER Tree

Different from other forwarding technologies, BIER does not need the transit nodes to maintain the multicast state, and only needs to forward and encapsulate the packets according to the BitString and BIFT. BIFT information can be learned by IGP. For each node of a sub-domain, the BIFT can be configured locally or allocated by BGP controller or PCE. Allocation by PCE is as per [I-D.chen-pce-pcep-extension-pce-controller-bier].

After receiving the BFR information reported from LHRs, the controller combine BFR information of LHRs into a BitString to guide multicast data forwarding.

The controller (PCE) sends the BitString to the FHR via PCInitiate or PCUpd message carrying Tree Forwarding State Synchronization (TFSS) object defined in Section 3.6 to inform the FHR to forward multicast data for a specific multicast tree according to the BitString, and the tree identifier is (*, g) or (s, g) tuple. Controller also needs to inform FHR to encapsulate VPN forwarding identifier so that LHRs can forward multicast data to specific VRF accordingly.

2.5. Multicast Statistics

Multicast statistics are helpful for multicast service analysis and business development. This section describes how to collect statistics about multicast users.

For each LHR, the statistics consist of two parts, one is the number of local users in the its managed domain, and the other is the number of other managed domains. The sum of these two parts is the number of multicast data that the one LHR needs to duplicate.

LHRs send this information to the controller via PCRpt message carrying Multicast Receiver Statistics (MRS) object. LHRs need report this information to the controller periodically. Once the statistics information of a LHR change, the LHR also needs reports to the controller.

Controller collects the statistics reported by LHRs and MAY periodically informs the FHR of the statistics via PCRpt messages carrying MRS object so that the FHR can feed the statistics back to the multicast source.

3. PCEP Object Formats

3.1. OPEN Object

3.1.1. STATEFUL-PCE-CAPABILITY TLV

During the PCEP initialization phase, PCEP speakers advertise stateful capability via the STATEFUL-PCE-CAPABILITY TLV in the OPEN object. Various flags are defined for the STATEFUL-PCE-CAPABILITY TLV defined in [RFC8231] and updated in [RFC8232] and [RFC8281].

A new flag is added in this document, whose code point is TBD1:

MULTICAST-STATE-CAPABILITY bit, if this bit is set to 1 by a PCEP speaker, it indicates that the PCEP speaker supports the capability of these new extensions as specified in this document.

To support the multicast tree state management capabilities described in this document, both PCE and PCC need to set MULTICAST-STATE-CAPABILITY bit. If a PCEP speaker receivers PECP message with the newly defined object, but without the MULTICAST-STATE-CAPABILITY bit set in STATEFUL-PCE-CAPABILITY TLV in the OPEN object, it MUST:

* Send a PCErr message with Error-Type=10(Reception of an invalid object) and Error-Value TBD2(MULTICAST-STATE-CAPABILITY bit is not set).

* Terminate the PCEP session.

3.1.2. PCECC Capability sub-TLV

The PCECC-CAPABILITY sub-TLV is an optional TLV for use in the OPEN Object for PCECC capability advertisement in PATH-SETUP-TYPE-CAPABILITY TLV as specified in [RFC9050].

[I-D.dhody-pce-pcep-extension-pce-controller-p2mp] adds a new flag (M Bit) in the PCECC-CAPABILITY sub-TLV to indicate the support for P2MP in PCECC, which is applicable for multicast tree described in this document as well.

3.2. PATH-SETUP-TYPE TLV

The PATH-SETUP-TYPE TLV is defined in [RFC8408]; [RFC9050] defines a PST value for PCECC as '2', which is applicable for multicast tree described in this document as well.

3.3. Multicast Source Registration Object

The MSR object is optional and SHOULD contain multicast source information and FHR information. The MSR object SHOULD be carried within a PCRpt message sent by PCC to PCE for registration. The MSR object SHOULD be carried within a PCUpd message sent by PCE to PCC in response to registration.

MSR Object-Class is TBD3. MSR Object-Type is 1. The format of the MSR object body is:

```
 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
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                            flags                            |R|A|
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|        Auxiliary Length       |            Reserved           |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
~                         Auxiliary Data                        ~
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
~                         Optional TLVs                         ~
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 1: MSR Object Body Format

R (Register flag, 1 bit): The R flag set to 1 indicates that the PCC is registering multicast information to the PCE. The R flag set to 0 indicates that the PCC revokes the registration.

A (Authentication flag, 1 bit): The A flag set to 1 indicates success of registration. The A flag set to 0 indicates failure of registration or revocation of registration. If there is no authentication information, A flag SHOULD also be set to 0.

Auxiliary Length (1 Octet): indicates the length of Auxiliary Data.

Auxiliary Data (Variable length): contains functional data such as authentication information.

MSR object MAY include Multicast Address TLV, mLDP FEC TLV, P2MP SR Policy Association Group Policy Identifiers TLV, and VPN Information TLV. P2MP SR Policy Association Group Policy Identifiers TLV are defined in [I-D.ietf-pce-sr-p2mp-policy]. MSR object carries different TLVs depending on the multicast tree.

3.3.1. Multicast Address TLV

In IP multicast scenarios, Multicast Address TLV SHOULD be included.

The format of the Multicast Address TLV is:

```
 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 = TBD4          |             Length            |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|         Prefix Length         |            Reserved           |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
~                     Multicast Source Address                  ~
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
~                     Multicast Group Address                   ~
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 2: Multicast Address TLV Format

Prefix Length (2 Octets): indicates the length of multicast addresses. the length MAY be 4 Octets, 8 Octets, 16 Octets, or 32 Octets. The multicast source address can be empty, but the multicast group address must be filled. If both the multicast source address and multicast group address exist, they must be both IPv4 and IPv6 addresses.

Multicast Source Address (Variable length): contains IPv4 or IPv6 address of the multicast source.

Multicast Group Address (Variable length): contains IPv4 or IPv6 address of the multicast group.

3.3.2. mLDP FEC TLV

In mLDP multicast scenarios, Multicast Address TLV SHOULD be included.

The format of the mLDP FEC TLV is:

```
 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          |             Length            |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
~                            mLDP FEC                           ~
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 3: mLDP FEC TLV Format

mLDP FEC (Variable length): carries mLDP FEC.

3.3.3. VPN Information TLV

VPN Information TLV is used to report VPN information about multicast sources and receivers. The format of the VPN Information TLV is:

```
 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          |             Length            |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                         VPN Identifier                        |
|                                                               |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 4: VPN Information TLV Format

VPN Identifier (8 Octets): indicates the VPN which the receiver used.

3.4. Multicast Receiver Information Object

The MRI object is optional and SHOULD contain receivers' information for matching the multicast registration information. The MRI object SHOULD be carried within a PCRpt message sent by PCC to PCE for multicast joining or leaving.

MRI Object-Class is TBD7. MRI Object-Type is 1. The format of the MRI object body is:

```
 0                   1                   2                   3
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|            Reserved           |            flags          |B|S|
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
~                        Optional TLVs                          ~
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 5: MRI Object Body Format

B(BIER multicast flag, 1 bit): The B flag set to 1 indicates that multicast protocol is BIER. If the B flag set to 1, MRI object SHOULD include BFR information TLV defined in Section 3.4.1.

S(Subscribe flag, 1 bit): The S flag set to 1 indicates that PCC delivers the message requesting to join PCE. The S flag set to 0 indicates that PCC delivers the message requesting to leave to PCE.

MRI object MAY include Multicast Address TLV, mLDP FEC TLV, P2MP SR Policy Association Group Policy Identifiers TLV, VPN Information TLV, and BFR information TLV.

3.4.1. BFR Information TLV

BFR Information TLV is used to report router location information in the BIER domain. The format of the BFR Information TLV is:

```
 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 = TBD8          |             Length            |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
| Subdomain-Id  |            BFR-ID             |  BSL  |Padding|
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 6: BFR Information TLV Format

Subdomain-Id (1 Octet): Unique value identifying the BIER subdomain.

BFR-ID (2 Octets): Identification of BFR in a subdomain.

BSL (BitString Length, 4 bits): encodes the length in bits of the BitString as per [RFC8296], the maximum length of the BitString is 7, it indicates the length of BitString is 4096. It is used to refer to the number of bits in the BitString.

3.5. CCI Object

The CCI object [RFC9050] is used by the PCE to specify the forwarding instructions to the PCC and MAY be carried within a PCInitiate or PCRpt message for control information download/report.

The CCI Object Type 1 for MPLS Label is defined in [RFC9050], which is used for the P2MP LSPs as well. For mLDP multicast trees, in addition to forwarding labels, corresponding tree labels or label stacks are required. Therefore, a new TLV is needed to carry tree labels or label stacks. In addition, in order to meet the needs of VPN scenarios, a new TLV with VPN forwarding Identifier is also defined.

3.5.1. Tree Label TLV

The format of Tree Label TLV is:

```
 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 = TBD9          |             Length            |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
~                        Tree Label stack                       ~
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 7: Tree Label TLV Format

Tree Label stack (Variable length): contains tree labels used to identify multicast trees. There are three application scenarios, as per [I-D.ietf-bess-bgp-multicast-controller].

3.5.2. VPN Forwarding Identifier TLV

The format of VPN Forwarding Identifier TLV is:

```
 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 = TBD10         |             Length            |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
~                   VPN Forwarding Identifier                   ~
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 8: VPN Forwarding Identifier TLV Format

VPN Forwarding Identifier (Variable length): contains MPLS label or IPv6 Segment Identifier with 16 Octets.

3.6. Tree Forwarding State Synchronization Object

The TFSS object is optional and SHOULD contain the forwarding state of control plane that the tree node needs to synchronize. The TFSS object SHOULD be carried within a PCInitiate or PCUpd message sent by PCE to PCC, or a PCRpt message sent by PCC to PCE for response.

TFSS Object-Class is TBD11. TFSS Object-Type is 1. The format of the TFSS object body is:

```
 0                   1                   2                   3
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|            Reserved           |              flags          |F|
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
~                        Optional TLVs                          ~
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 9: TFSS Object Body Format

F (Forwarding flag, 1 bit): The F flag set to 1 indicates that the node MAY start forwarding multicast packets. The F flag set to 0 indicates that the node SHOULD stop forwarding multicast packets.

TFSS object MAY include Multicast Address TLV, VPN Forwarding Identifier TLV, and BIER Attribute TLV. When TT=1, BIER Attribute TLV SHOULD be included.

3.6.1. BIER Attribute TLV

BIER Attribute TLV is used to instruct the FHR to forward multicast packets to the users according to the BitString specified by the controller. The format of BIER Attribute TLV is:

```
 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 = TBD12         |             Length            |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
| Subdomain-id  |       SI      |  BSL  ~       BitString       ~
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 10: BIER Attribute TLV Format

Subdomain-id (1 Octet): Unique value identifying the BIER subdomain.

SI (Set Identifier, 1 Octet): encoding the Set Identifier used in the encapsulation for this BIER subdomain for this BitString length.

BSL (BitString Length, 4 bits): encodes the length in bits of the BitString as per [RFC8296], the maximum length of the BitString is 7, it indicates the length of BitString is 4096. It is used to refer to the number of bits in the BitString.

BitString(Variable length): indicates the path of multicast data packets forwarding for headend.

3.7. Multicast Receiver Statistics Object

The MRS object is optional and used to inform PCE of the number of receivers. The MRS object SHOULD be carried within a PCRpt or a PCUpd message for synchronize receiver information periodically, or PCRpt message for the leaving of receivers.

MRS Object-Class is TBD13. MRS Object-Type is 1. The format of the MRS object body is:

```
 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
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|          Number Length        |            Reserved           |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                      Multicast Statistics                     |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
~                        Optional TLVs                          ~
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 11: MRS Object Body Format

Number Length (2 Octets): indicates the length of receiver number.

Multicast Statistics (4 Octets): indicates the statistics information for a particular multicast tree of a controller or a LHR.

MRS object MAY include Multicast Address TLV, mLDP FEC TLV, P2MP SR Policy Association Group Policy Identifiers TLV.

4. Message Formats

4.1. PCRpt message

The definition of the PCRpt message from [RFC8231] and [RFC9050] is extended to optionally include MSR object, MRI object, TFSS object, and MRS object after the path object. The encoding will become:

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

Where:

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

<state-report> ::= [<SRP>]
                   <LSP>
                   <path>
                   [<MSR>]
                   [<MRI>]
                   [<TFSS>]
                   [<MRS>]
```

Where:
<path> is as per [RFC8231] and the LSP and SRP object are also defined in [RFC8231].

4.2. PCInitiate message

The definition of the PCInitiate message from [RFC8281] is extended to optionally include TFSS object after the path object. The encoding will become:

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

Where:

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

<update-request> ::= <SRP>
                     <LSP>
                     [<TFSS>]
```

Where:
LSP and SRP object are defined in [RFC8231].

4.3. PCUpd message

The definition of the PCUpd message from [RFC8231] is extended to optionally include MSR object, TFSS object and MRS object after the path object. The encoding will become:

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

Where:

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

<update-request> ::= <SRP>
                     <LSP>
                     <path>
                     [<MSR>]
                     [<TFSS>]
                     [<MRS>]
```

Where:
<path> is as per [RFC8231] and the LSP and SRP object are also defined in [RFC8231].

5. Example Workflow

5.1. Multicast Tree Discovery

```
                 +-------+                      +-------+
                 |PCC    |                      |  PCE  |
                 |ingress|                      +-------+
          +-----|       |                          |
          |PCC  +-------+                          |
          |transit| |                              |
    +------|       | |                              |
    |PCC   +-------+ |                              |
    |egress |  |     |                              |
    +-------+  |     |                              |
        |      |     |-----PCRpt,PLSP-ID=1,D=1------>|Source Registration
        |      |     |       MSR Object(R=1)        |(Root Discovery)
        |      |     |                              |
        |      |     |<----PCUpd,PLSP-ID=1-----------|Source Registration
        |      |     |     MSR Object(R=1)          |Confirmation
        |      |     |                              |
        |------------------PCRpt,PLSP-ID=0---------->|Receiver Report
        |      |     |     MRI Object(S=1)          |(Leaf Discovery)
        |      |     |                              |
```

Figure 11: Workflow of Multicast Tree Discovery

5.2. Multicast Tree Path Establishment and Update

5.2.1. Labeled Tree

The workflow of Labeled Tree are as per [I-D.ietf-pce-sr-p2mp-policy].

5.2.2. BIER Tree

```
                  +-------+                        +-------+
                  |PCC    |                        |  PCE  |
                  |ingress|                        +-------+
          +------|       |                            |
          |      +-------+                            |
          | transit| |                               |
    +------|        | |                               |
    |      +--------+ |                               |
    |egress  |  |     |                               |
    +--------+  |     |                               |
        |<--------------------PCInitiate,PLSP-ID=1-------|Tree State
        |       |     |           TFSS Object           |Setup
        |       |     |     VPN Forwarding Identifier   |
        |       |     |                                 |
        |--------------------PCRpt,PLSP-ID=1----------->|Tree State
        |       |     |                                 |Confirmation
        |       |     |                                 |
        |       |     |<-----PCInitiate,PLSP-ID=1-------|Tree State
        |       |     |      TFSS Object (F=1,TT=1)     |Setup
        |       |     |     VPN Forwarding Identifier   |
        |       |     |                                 |
        |       |     |------PCRpt,PLSP-ID=1----------->|Tree State
        |       |     |                                 |Confirmation
        |       |     |                                 |
        |       |     |                                 |
        |       |     |                                 |
        |--------------------PCRpt,PLSP-ID=0----------->|Receiver Report
        |       |     |      MRI Object(S=1/S=0)        |(Leaf Discovery)
        |       |     |                                 |
        |<--------------------PCInitiate,PLSP-ID=1-------|Tree State
        |       |     |           TFSS Object           |Update
        |       |     |     VPN Forwarding Identifier   |
        |       |     |                                 |
        |--------------------PCRpt,PLSP-ID=1----------->|Tree State
        |       |     |                                 |Confirmation
        |       |     |                                 |
        |       |     |<-----PCUpd,PLSP-ID=1------------|Tree State
        |       |     |      TFSS Object (F=1,TT=1)     |Update
        |       |     |                                 |
        |       |     |------PCRpt,PLSP-ID=1----------->|Tree State
        |       |     |                                 |Confirmation
        |       |     |                                 |
```

Figure 12: Workflow of BIER Tree Setup/Update

In the previous multicast tree discovery process, ingress has delegated the PCE to calculate the multicast path. Therefore, during the establishment of the multicast tree, PCE informs the ingress to forward multicast data for (S,G)/(*,G) via PCInitiate message carrying TFSS object according to the delegation. The PLSP-ID is the value specified by the ingress. PCE also sends VPN forwarding identifier to ingress and egresses, so that multicast packets can be forwarded to specified VPN.

Ingress then updates the tree state and responds to the PCE. At this point, the BIER tree is formally established and the ingress starts forwarding for the multicast according to the BitString specified by PCE.

If the topology of the BIER tree changes, the corresponding egresses will send PCRpt messages carring MRI object to PCE to inform PCE of their changes. PCE needs to send VPN forwarding identifier to these egresses.

The PCE updates the BitStrig based on changes of the egresses and informs the ingress to update. State update process is similar to state setup process, except for the type of message sent by the PCE.

5.3. Multicast Statistics Synchronization

```
                  +-------+                       +-------+
                  |PCC    |                       |  PCE  |
                  |ingress|                       +-------+
           +-----|       |                           |
           |PCC  +-------+                           |
           |transit| |                               |
     +-----|       | |                               |
     |PCC   +-------+ |                               |
     |egress |  |     |                               |
     +-------+  |     |                               |
         |      |     |                               |
         |-----------------PCRpt,PLSP-ID=0---------->| Statistics
         |      |     |        MRS Object             | Report
         |      |     |case1: Periodic Synchronization|
         |      |     | case2: Change Synchronization |
         |      |     |                               |
         |      |     |                               |
         |      |     |<----PCUpd,PLSP-ID=0-----------| Statistics
         |      |     |       MRS Object              | Feedback
         |      |     |   Periodic Synchronization    |
         |      |     |                               |
```

Figure 13: Workflow of Multicast Statistics Synchronization

6. Security Considerations

To be added.

7. IANA Considerations

7.1. MULTICAST-STATE-CAPABILITY

IANA is requested to allocate a new code point within registry "STATEFUL-PCE-CAPABILITY TLV Flag Field" under "Path Computation Element Protocol (PCEP) Numbers" as follows:

Value	Description	Reference
TBD1	MULTICAST-STATE-CAPABILITY	This document

7.2. PCEP-ERROR Object

IANA is requested to allocate code-points in the "PCEP-ERROR Object Error Types and Values" sub-registry for the following new error-type and error-value:

Error-Type	Description	Reference
10	Error-value = TBD2 MULTICAST-STATE-CAPABILITY bit is not set	This document

7.3. New Objects

IANA is requested to allocate the following Object-Class Values in the "PCEP Objects" sub-registry under the "Path Computation Element Protocol (PCEP) Numbers" registry:

Object-Class Value	Description	Reference
TBD3	Multicast Source Registration	This document
TBD7	Multicast Receiver Information	This document
TBD11	Tree Forwarding State Synchronization	This document
TBD13	Multicast Receiver Statistics	This document

IANA is requested to allocate the following Object-Type Value of CCI object in the "PCEP Objects" sub-registry under the "Path Computation Element Protocol (PCEP) Numbers" registry:

7.4. New TLVs

IANA is requested to allocate the following TLVs:

Type	Description	Reference
TBD4	Multicast Source Address	This document
TBD5	mLDP FEC	This document
TBD6	VPN Information	This document
TBD8	BFR Information	This document
TBD9	Tree Label	This document
TBD10	VPN Forwarding Identifier TLV	This document
TBD12	BIER Attribute	This document

8. Acknowledgements

The author thanks ... for their review and comments.

9. References

9.1. Normative References

[I-D.ietf-pce-sr-p2mp-policy]
Bidgoli, H., Voyer, D., Rajarathinam, S., Hemmati, E., Saad, T., and S. Sivabalan, "PCEP extensions for p2mp sr policy", Work in Progress, Internet-Draft, draft-ietf-pce-sr-p2mp-policy-01, 7 July 2022, <https://www.ietf.org/archive/id/draft-ietf-pce-sr-p2mp-policy-01.txt>.

[RFC1112] Deering, S., "Host extensions for IP multicasting", STD 5, RFC 1112, DOI 10.17487/RFC1112, August 1989, <https://www.rfc-editor.org/info/rfc1112>.

[RFC2362] Estrin, D., Farinacci, D., Helmy, A., Thaler, D., Deering, S., Handley, M., Jacobson, V., Liu, C., Sharma, P., and L. Wei, "Protocol Independent Multicast-Sparse Mode (PIM-SM): Protocol Specification", RFC 2362, DOI 10.17487/RFC2362, June 1998, <https://www.rfc-editor.org/info/rfc2362>.

[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>.

[RFC4875] Aggarwal, R., Ed., Papadimitriou, D., Ed., and S. Yasukawa, Ed., "Extensions to Resource Reservation Protocol - Traffic Engineering (RSVP-TE) for Point-to-

Multipoint TE Label Switched Paths (LSPs)", RFC 4875, DOI 10.17487/RFC4875, May 2007, <https://www.rfc-editor.org/info/rfc4875>.

[RFC5015] Handley, M., Kouvelas, I., Speakman, T., and L. Vicisano, "Bidirectional Protocol Independent Multicast (BIDIR-PIM)", RFC 5015, DOI 10.17487/RFC5015, October 2007, <https://www.rfc-editor.org/info/rfc5015>.

[RFC6388] Wijnands, IJ., Ed., Minei, I., Ed., Kompella, K., and B. Thomas, "Label Distribution Protocol Extensions for Point-to-Multipoint and Multipoint-to-Multipoint Label Switched Paths", RFC 6388, DOI 10.17487/RFC6388, November 2011, <https://www.rfc-editor.org/info/rfc6388>.

[RFC6514] Aggarwal, R., Rosen, E., Morin, T., and Y. Rekhter, "BGP Encodings and Procedures for Multicast in MPLS/BGP IP VPNs", RFC 6514, DOI 10.17487/RFC6514, February 2012, <https://www.rfc-editor.org/info/rfc6514>.

[RFC7761] Fenner, B., Handley, M., Holbrook, H., Kouvelas, I., Parekh, R., Zhang, Z., and L. Zheng, "Protocol Independent Multicast - Sparse Mode (PIM-SM): Protocol Specification (Revised)", STD 83, RFC 7761, DOI 10.17487/RFC7761, March 2016, <https://www.rfc-editor.org/info/rfc7761>.

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

[RFC8232] Crabbe, E., Minei, I., Medved, J., Varga, R., Zhang, X., and D. Dhody, "Optimizations of Label Switched Path State Synchronization Procedures for a Stateful PCE", RFC 8232, DOI 10.17487/RFC8232, September 2017, <https://www.rfc-editor.org/info/rfc8232>.

[RFC8279] Wijnands, IJ., Ed., Rosen, E., Ed., Dolganow, A., Przygienda, T., and S. Aldrin, "Multicast Using Bit Index Explicit Replication (BIER)", RFC 8279, DOI 10.17487/RFC8279, November 2017, <https://www.rfc-editor.org/info/rfc8279>.

[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>.

[RFC8296] Wijnands, IJ., Ed., Rosen, E., Ed., Dolganow, A., Tantsura, J., Aldrin, S., and I. Meilik, "Encapsulation for Bit Index Explicit Replication (BIER) in MPLS and Non-MPLS Networks", RFC 8296, DOI 10.17487/RFC8296, January 2018, <https://www.rfc-editor.org/info/rfc8296>.

[RFC8402] Filsfils, C., Ed., Previdi, S., Ed., Ginsberg, L., Decraene, B., Litkowski, S., and R. Shakir, "Segment Routing Architecture", RFC 8402, DOI 10.17487/RFC8402, July 2018, <https://www.rfc-editor.org/info/rfc8402>.

[RFC9050] Li, Z., Peng, S., Negi, M., Zhao, Q., and C. Zhou, "Path Computation Element Communication Protocol (PCEP) Procedures and Extensions for Using the PCE as a Central Controller (PCECC) of LSPs", RFC 9050, DOI 10.17487/RFC9050, July 2021, <https://www.rfc-editor.org/info/rfc9050>.

9.2. Informative References

[I-D.chen-pce-pcep-extension-pce-controller-bier] Chen, R., Xu, B., Chen, H., and A. Wang, "PCEP Procedures and Protocol Extensions for Using PCE as a Central Controller (PCECC) of BIER", Work in Progress, Internet-Draft, draft-chen-pce-pcep-extension-pce-controller-bier-03, 7 March 2022, <https://www.ietf.org/archive/id/draft-chen-pce-pcep-extension-pce-controller-bier-03.txt>.

[I-D.dhody-pce-pcep-extension-pce-controller-p2mp] Li, Z., Peng, S., Geng, X., and M. S. Negi, "Path Computation Element Communication Protocol (PCEP) Extensions for Using the PCE as a Central Controller (PCECC) of point-to-multipoint (P2MP) LSPs", Work in Progress, Internet-Draft, draft-dhody-pce-pcep-extension-pce-controller-p2mp-09, 10 July 2022, <https://www.ietf.org/archive/id/draft-dhody-pce-pcep-extension-pce-controller-p2mp-09.txt>.

[I-D.ietf-bess-bgp-multicast-controller] Zhang, Z., Raszuk, R., Pacella, D., and A. Gulko, "Controller Based BGP Multicast Signaling", Work in Progress, Internet-Draft, draft-ietf-bess-bgp-multicast-

controller-09, 11 April 2022, <https://www.ietf.org/archive/id/draft-ietf-bess-bgp-multicast-controller-09.txt>.

[I-D.ietf-pim-sr-p2mp-policy]
(editor), D. V., Filsfils, C., Parekh, R., Bidgoli, H., and Z. Zhang, "Segment Routing Point-to-Multipoint Policy", Work in Progress, Internet-Draft, draft-ietf-pim-sr-p2mp-policy-05, 2 July 2022, <https://www.ietf.org/archive/id/draft-ietf-pim-sr-p2mp-policy-05.txt>.

[I-D.ietf-spring-sr-replication-segment]
(editor), D. V., Filsfils, C., Parekh, R., Bidgoli, H., and Z. Zhang, "SR Replication Segment for Multi-point Service Delivery", Work in Progress, Internet-Draft, draft-ietf-spring-sr-replication-segment-08, 1 July 2022, <https://www.ietf.org/archive/id/draft-ietf-spring-sr-replication-segment-08.txt>.

[I-D.zzhang-mvpn-evpn-controller]
Zhang, Z., Parekh, R., Zhang, Z., and H. Bidgoli, "MVPN and EVPN BUM Signaling with Controllers", Work in Progress, Internet-Draft, draft-zzhang-mvpn-evpn-controller-01, 25 October 2021, <https://www.ietf.org/archive/id/draft-zzhang-mvpn-evpn-controller-01.txt>.

Authors' Addresses

Huanan Li
China Telecom
Email: lihn6@foxmail.com

Aijun Wang
China Telecom
Email: wangaj3@chinatelecom.cn

Zhaohui Zhang
Juniper Networks
Email: zzhang@juniper.net

Huaimo Chen
Futurewei
Email: Huaimo.chen@futurewei.com

Ran Chen
ZTE Corporation
Email: chen.ran@zte.com.cn

PCE
Internet-Draft
Intended status: Standards Track
Expires: 3 September 2022

Q. Xiong
S. Peng
ZTE Corporation
F. Qin
China Mobile
March 2022

PCEP Extension for SR-MPLS Entropy Label Position
draft-peng-pce-entropy-label-position-07

Abstract

This document proposes a set of extensions for PCEP to configure the entropy label position for SR-MPLS networks.

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 2 September 2022.

Table of Contents

1. Introduction

[RFC5440] describes the Path Computation Element Protocol (PCEP) which is used between a Path Computation Element (PCE) and a Path Computation Client (PCC) (or other PCE) to enable computation of Multi-protocol Label Switching (MPLS) for Traffic Engineering Label Switched Path (TE LSP). PCEP Extensions for the Stateful PCE Model [RFC8231] describes a set of extensions to PCEP to enable active control of MPLS-TE and Generalized MPLS (GMPLS) tunnels. [RFC8281] describes the setup and teardown of PCE-initiated LSPs under the active stateful PCE model, without the need for local configuration on the PCC, thus allowing for dynamic centralized control of a network.

Segment Routing (SR) leverages the source routing paradigm. Segment Routing can be instantiated on MPLS data plane which is referred to as SR-MPLS [RFC8660]. SR-MPLS leverages the MPLS label stack to construct the SR path. PCEP Extensions for Segment Routing [RFC8664] specifies extensions to the PCEP that allow a stateful PCE to compute and initiate TE paths, as well as a PCC to request a path subject to certain constraint(s) and optimization criteria in SR networks.

Entropy label (EL) [RFC6790] is a technique used in the MPLS data plane to improve load-balancing. Entropy Label Indicator (ELI) can be immediately preceding an EL in the MPLS label stack. The idea behind the EL is that the ingress router computes a hash based on several fields from a given packet and places the result in an

additional label, named "entropy label". Then, this entropy label can be used as part of the hash keys used by an LSR. Using the entropy label as part of the hash keys reduces the need for deep packet inspection in the LSR while keeping a good level of entropy in the load-balancing. When the entropy label is used, the keys used in the hashing functions are still a local configuration matter and an LSR may use solely the entropy label or a combination of multiple fields from the incoming packet.

[RFC8662] proposes to use entropy labels for SR-MPLS networks and mutiple <ELI, EL> pairs SHOULD be inserted in the SR-MPLS label stack. The ingress node may decide the number and place of the ELI/ELs which need to be inserted into the label stack. The extensions for Border Gateway Protocol (BGP) to indicate the entropy label position in the SR-MPLS label stack has been proposed in [I-D.zhou-idr-bgp-srmpls-elp].

In some cases, the the controller(e.g. PCE) could be used to perform the TE path computation as well as the Entropy Label Position (ELP) which is useful for inter-domain scenarios. This document proposes a set of extensions for PCEP to configure the ELP information for SR-MPLS networks.

2. Conventions used in this document

2.1. Terminology

The terminology is defined as [RFC5440], [RFC6790], [RFC8664] and [RFC8662].

2.2. Requirements Language

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

3. Entropy Labels in SR-MPLS Scenario with PCE

[RFC8662] proposes to use entropy labels for SR-MPLS networks. The Entropy Readable Label Depth (ERLD) is defined as the number of labels which means that the router will perform load-balancing using the ELI/EL. An appropriate algorithm should consider the following criteria:

* a limited number of <ELI, EL> pairs SHOULD be inserted in the SR-MPLS label stack;

* the inserted positions SHOULD be whithin the ERLD of a maximize number of transit LSRs;

* a minimum number of <ELI, EL> pairs SHOULD be inserted while satisfying the above criteria.

As described in [RFC8662] section 7, the ERLD value is important for inserting ELI/EL and the ingress node need to evaluate the minimum ERLD value along the node segment path. But it will add complexity in the ELI/EL insertion process. Moreover, the ingress node cannot find the minimum ERLD along the path and does not support the computation of the minimum ERLD especilly in inter-domain scenarios. As the Figure 1 shown, in SR-MPLS inter-domain scenario, the ingress node of the first domain could not get the ERLD information of other nodes of other domains.

```
        +-----+                +-----+                +-----+
        |PCE-1|                |PCE-2|                |PCE-3|
        +--+--+                +--+--+                +--+--+
           |                      |                      |
  .........+.........   .........+.........    .........+..........
  .                 .   .                 .    .                  .
  .+---+      +---+ .   . +---+     +---+ .    .+---+     +----+   .
  .| A |------| B |------ | C |-----| X |------| Y |-----| Z  |   .
  .+---+      +---+ .   . +---+     +---+ .    .+---+     +----+   .
  .    SR-AS 1      .   .   SR-AS 2       .    .     SR-AS 3      .
  ....................   ....................    ......................
```

Figure 1: Figure 1: Entropy Labels in SR-MPLS Inter-Domain Scenario

The PCEs could get the information of all nodes such as Maximum SID Depth (MSD) and ERLD through Interior Gateway Protocol (IGP) and can compute the minimum ERLD along the end-to-end path. For example, the ERLD value can be collected via IS-IS [I-D.ietf-isis-mpls-elc], OSPF[I-D.ietf-ospf-mpls-elc]. [RFC8476] and [RFC8491] provide examples of advertisement of the MSD. Moreover, the PCEs also can compute the Entropy Label Position (ELP) including the number and the places of the ELI/ELs. Then the ingress nodes MAY be required to support the capabilities of inserting multiple ELI/ELs and need to advertise the capabilities to the PCEs.

This document proposes the extensions for PCE to perform the computation of the end-to-end path as well as the positions of entropy labels in SR-MPLS networks. The ingress nodes can directly insert the ELI/ELs based on the positions.

4. PCEP Extensions

4.1. The OPEN Object

As defined in [RFC8664], PCEP speakers use SR PCE Capability sub-TLV to exchange information about their SR capability when PST=1 in the PST List of the PATH-SETUP-TYPE-CAPABILITY TLV carried in Open object. This document defined a new flag (E-flag) for SR PCE Capability sub-TLV as shown in Figure 2.

```
 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=TBD11            |            Length=4           |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|         Reserved              |   Flags |E|N|X|      MSD      |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 2: Figure 2: E-flag in SR-PCE-CAPABILITY sub-TLV

E (Entropy Label Configuration is supported) : A PCE sets this flag bit to 1 carried in Open message to indicate that it supports the computation of SR path with ELP information. A PCC sets this flag to 1 to indicate that it supports the capability of inserting multiple ELI/EL pairs and and supports the results of SR path with ELP from PCE.

4.2. The LSP-EXTENDED-FLAG TLV

The LSP Object is defined in Section 7.3 of [RFC8231]. This document defiend a new flag (E-flag) for the LSP-EXTENDED-FLAG TLV carried in LSP Object as defined in [I-D.ietf-pce-lsp-extended-flags]. The format is shown as Figure 3:

```
 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=TBD            |          Length               |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                                                             |E|
//                 LSP Extended Flags                          //
|                                                               |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 3: Figure 3: E-flag in LSP-EXTENDED-FLAG TLV

E (Request for ELP Configuration) : If the bit is set to 1, it indicates that the PCC requests PCE to compute the SR path with ELP information. A PCE would also set this bit to 1 to indicate that the ELP information is included by PCE and encoded in the PCRep, PCUpd or PCInitiate message.

4.3. The SR-ERO Object

SR-ERO subobject is used for SR-TE path which consists of one or more SIDs as defined in [RFC8664]. This document defiend a new flag (E-flag) for the SR-ERO subobject as Figure 4 shown:

```
 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
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|L|   Type=36   |     Length    |  NT   |     Flags   |E|F|S|C|M|
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                         SID (optional)                        |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
//                   NAI (variable, optional)                  //
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 4: Figure 4: E-flag in SR-ERO subobject

E (ELP Configuration) : If this flag is set, it means that the position after this SR-ERO subobject is the position to insert <ELI, EL>, otherwise it cannot insert <ELI, EL> after this segment.

5. Operations

The SR path is initiated by PCE or PCC with PCReq, PCInitiated or PCUpd messages and the E bit is set to 1 in LSP object to request the ELP configuration. The SR-TE path being recieved by PCC with SR-ERO segment list, for example, <S1, S2, S3, S4, S5, S6>, especially S3 and S6 with E-flag set. It indicates that two <ELI, EL> pairs MUST be inserted into the label stack of the SR-TE forwarding entry, repectively after the label for S3 and label for S6. With EL information, the label stack for SR-MPLS would be <label1, label2, label3, ELI, EL, label4, label5, label6, ELI, EL>.

6. Security Considerations

Procedures and protocol extensions defined in this document do not introduce any new security considerations beyond those already listed in [RFC8662] and [RFC8664].

7. Acknowledgements

The authors would like to thank Stephane Litkowski, Dhruv Dhody, Tarek Saad, Zhenbin Li and Jeff Tantsura for their review, suggestions and comments to this document.

8. IANA Considerations

8.1. New SR PCE Capability Flag Registry

SR PCE Capability TLV is defined in [RFC8664], and the registry to manage the Flag field of the SR PCE Capability TLV is requested in [RFC8664]. IANA is requested to make allocations from the registry, as follows:

Value	Name	Reference
TBD11	Entropy Label Configuration is supported (E)	[this document]

Table 1

8.2. New LSP-EXTENDED-FLAG Flag Registry

[I-D.ietf-pce-lsp-extended-flags] defines the LSP-EXTENDED-FLAG TLV. IANA is requested to make allocations from the Flag field registry, as follows:

Value	Name	Reference
TBD	Request for ELP Configuration (E)	[this document]

Table 2

8.3. New SR-ERO Flag Registry

SR-ERO subobject is defined in [RFC8664], and the registry to manage the Flag field of SR-ERO is requested in [RFC8664]. IANA is requested to make allocations from the registry, as follows:

Value	Name	Reference
36	ELP Configuration (E)	[this document]

Table 3

9. Normative References

[I-D.ietf-isis-mpls-elc]
Xu, X., Kini, S., Psenak, P., Filsfils, C., Litkowski, S., and M. Bocci, "Signaling Entropy Label Capability and Entropy Readable Label Depth Using IS-IS", Work in Progress, Internet-Draft, draft-ietf-isis-mpls-elc-13, 28 May 2020, <https://www.ietf.org/archive/id/draft-ietf-isis-mpls-elc-13.txt>.

[I-D.ietf-ospf-mpls-elc]
Xu, X., Kini, S., Psenak, P., Filsfils, C., Litkowski, S., and M. Bocci, "Signaling Entropy Label Capability and Entropy Readable Label Depth Using OSPF", Work in Progress, Internet-Draft, draft-ietf-ospf-mpls-elc-15, 1 June 2020, <https://www.ietf.org/archive/id/draft-ietf-ospf-mpls-elc-15.txt>.

[I-D.ietf-pce-lsp-extended-flags]
Xiong, Q., "LSP Object Flag Extension of Stateful PCE", Work in Progress, Internet-Draft, draft-ietf-pce-lsp-extended-flags-01, 18 October 2021, <https://www.ietf.org/archive/id/draft-ietf-pce-lsp-extended-flags-01.txt>.

[I-D.zhou-idr-bgp-srmpls-elp]
Liu, Y. and S. Peng, "BGP Extension for SR-MPLS Entropy Label Position", Work in Progress, Internet-Draft, draft-zhou-idr-bgp-srmpls-elp-04, 1 March 2022, <https://www.ietf.org/archive/id/draft-zhou-idr-bgp-srmpls-elp-04.txt>.

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <https://www.rfc-editor.org/info/rfc2119>.

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

[RFC6790] Kompella, K., Drake, J., Amante, S., Henderickx, W., and L. Yong, "The Use of Entropy Labels in MPLS Forwarding", RFC 6790, DOI 10.17487/RFC6790, November 2012, <https://www.rfc-editor.org/info/rfc6790>.

[RFC8174] Leiba, B., "Ambiguity of Uppercase vs Lowercase in RFC 2119 Key Words", BCP 14, RFC 8174, DOI 10.17487/RFC8174, May 2017, <https://www.rfc-editor.org/info/rfc8174>.

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

[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>.

[RFC8476] Tantsura, J., Chunduri, U., Aldrin, S., and P. Psenak, "Signaling Maximum SID Depth (MSD) Using OSPF", RFC 8476, DOI 10.17487/RFC8476, December 2018, <https://www.rfc-editor.org/info/rfc8476>.

[RFC8491] Tantsura, J., Chunduri, U., Aldrin, S., and L. Ginsberg, "Signaling Maximum SID Depth (MSD) Using IS-IS", RFC 8491, DOI 10.17487/RFC8491, November 2018, <https://www.rfc-editor.org/info/rfc8491>.

[RFC8623] Palle, U., Dhody, D., Tanaka, Y., and V. Beeram, "Stateful Path Computation Element (PCE) Protocol Extensions for Usage with Point-to-Multipoint TE Label Switched Paths (LSPs)", RFC 8623, DOI 10.17487/RFC8623, June 2019, <https://www.rfc-editor.org/info/rfc8623>.

[RFC8660] Bashandy, A., Ed., Filsfils, C., Ed., Previdi, S., Decraene, B., Litkowski, S., and R. Shakir, "Segment Routing with the MPLS Data Plane", RFC 8660, DOI 10.17487/RFC8660, December 2019, <https://www.rfc-editor.org/info/rfc8660>.

[RFC8662] Kini, S., Kompella, K., Sivabalan, S., Litkowski, S., Shakir, R., and J. Tantsura, "Entropy Label for Source Packet Routing in Networking (SPRING) Tunnels", RFC 8662, DOI 10.17487/RFC8662, December 2019, <https://www.rfc-editor.org/info/rfc8662>.

[RFC8664] Sivabalan, S., Filsfils, C., Tantsura, J., Henderickx, W., and J. Hardwick, "Path Computation Element Communication Protocol (PCEP) Extensions for Segment Routing", RFC 8664, DOI 10.17487/RFC8664, December 2019, <https://www.rfc-editor.org/info/rfc8664>.

Authors' Addresses

Quan Xiong
ZTE Corporation
No.6 Huashi Park Rd
Wuhan
Hubei, 430223
China
Email: xiong.quan@zte.com.cn

Shaofu Peng
ZTE Corporation
No.50 Software Avenue
Nanjing
Jiangsu, 210012
China
Email: peng.shaofu@zte.com.cn

Fengwei Qin
China Mobile
Beijing
China

Email: qinfengwei@chinamobile.com

PCE
Internet-Draft
Intended status: Standards Track
Expires: 9 August 2024

Q. Xiong
S. Peng
ZTE Corporation
F. Qin
China Mobile
J. Zhao
CAICT

6 February 2024

Path Computation Element Communication Protocol (PCEP) Extension for SR-MPLS Entropy Label Positions
draft-peng-pce-entropy-label-position-11

Abstract

Entropy label (EL) can be used in the SR-MPLS data plane to improve load-balancing and multiple Entropy Label Indicator (ELI)/EL pairs SHOULD be inserted in the SR-MPLS label stack as per RFC8662.

This document proposes a set of extensions for Path Computation Element Communication Protocol (PCEP) to configure the Entropy Label Positions (ELP) for SR-MPLS networks.

Status of This Memo

Copyright Notice

Table of Contents

1. Introduction

[RFC5440] describes the Path Computation Element Computation Protocol (PCEP) which is used between a Path Computation Element (PCE) and a Path Computation Client (PCC) (or other PCE) to enable computation of Multi-protocol Label Switching (MPLS) for Traffic Engineering Label Switched Path (TE LSP). PCEP Extensions for the Stateful PCE Model [RFC8231] describes a set of extensions to PCEP to enable active control of MPLS-TE and Generalized MPLS (GMPLS) tunnels. [RFC8281] describes the setup and teardown of PCE-initiated LSPs under the active stateful PCE model, without the need for local configuration on the PCC, thus allowing for dynamic centralized control of a network.

Segment Routing (SR) leverages the source routing paradigm. Segment Routing can be instantiated on MPLS data plane which is referred to as SR-MPLS [RFC8660]. SR-MPLS leverages the MPLS label stack to construct the SR path. PCEP Extensions for Segment Routing [RFC8664] specifies extensions to the PCEP that allow a stateful PCE to compute and initiate TE paths, as well as a PCC to request a path subject to certain constraint(s) and optimization criteria in SR networks.

Entropy label (EL) [RFC6790] is a technique used in the MPLS data plane to improve load-balancing. Entropy Label Indicator (ELI) can be immediately preceding an EL in the MPLS label stack. The idea behind the EL is that the ingress router computes a hash based on several fields from a given packet and places the result in an additional label, named "entropy label". Then, this entropy label can be used as part of the hash keys used by an Label Switch Router (LSR). Using the entropy label as part of the hash keys reduces the need for deep packet inspection in the LSR while keeping a good level of entropy in the load-balancing. When the entropy label is used, the keys used in the hashing functions are still a local configuration matter and an LSR may use solely the entropy label or a combination of multiple fields from the incoming packet.

[RFC8662] proposes to use entropy labels for SR-MPLS networks and multiple <ELI, EL> pairs SHOULD be inserted in the SR-MPLS label stack. The ingress node may decide the number and place of the ELI/ELs which need to be inserted into the label stack. The Entropy Label Position (ELP) is used to indicate the positions of the ELI/ELs which need to be inserted into the label stack as per [I-D.ietf-idr-bgp-srmpls-elp].

In some cases, the the controller(e.g. PCE) could be used to perform the TE path computation as well as ELP information which is useful for inter-domain scenarios. This document proposes a set of extensions for PCEP to configure the ELP information for SR-MPLS networks.

2. Conventions used in this document

2.1. Terminology

The terminology is defined as [RFC5440], [RFC6790], [RFC8664] and [RFC8662].

2.2. Requirements Language

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

3. Entropy Labels in SR-MPLS Scenario with PCE

[RFC8662] proposes to use entropy labels for SR-MPLS networks. The Entropy Readable Label Depth (ERLD) is defined as the number of labels which means that the router will perform load-balancing using the ELI/EL in [RFC8662] section 4.

As described in [RFC8662] section 7.2.1, the ELRD value is an important consideration when inserting ELI/EL and the minimum ELRD must be evaluated for each node along a computed path. This necessary step adds additional complexity in the ELI/EL insertion process and it may not be feasible for an ingress router to compute the appropriate ERLD for each node in the path, since a SR-MPLS path may contain segments the ingress router can resolve such as inter-domain scenarios. As the Figure 1 shown, in SR-MPLS inter-domain scenario, the ingress node of the first domain could not get the ERLD information of other nodes of other domains.

```
       +-----+                +-----+                 +-----+
       |PCE-1|                |PCE-2|                 |PCE-3|
       +--+--+                +--+--+                 +--+--+
          |                      |                       |
 .........+.........   .........+.........    .........+...........
 .                 .   .                 .    .                   .
 .+---+       +---+ .   . +---+      +---+ .    .+---+      +----+  .
 .| A |-------| B |------ | C |------| X |-------| Y |------| Z  |  .
 .+---+       +---+ .   . +---+      +---+ .    .+---+      +----+  .
 .    SR-AS 1       .   .   SR-AS 2        .    .     SR-AS 3       .
 ....................   ....................    .....................
```

Figure 1: Entropy Labels in SR-MPLS Inter-Domain Scenario

When computing the ELI/EL positions, the PCE MUST take into consideration Maximum SID Depth (MSD) imposition. The PCEs could get the information of all nodes such as MSD (e.g. Base MPLS Imposition MSD (BMI-MSD) or ERLD-MSD) through Interior Gateway Protocol (IGP) and can compute the minimum ERLD along the end-to-end path. IS-IS [RFC8491] and OSPF [RFC8476] provide examples of advertisement of the MSD. The ERLD value can be collected via IS-IS [RFC9088], and OSPF [RFC9089]. Moreover, the PCEs also can compute the ELP information including the number and the places of the ELI/ELs. Then the ingress nodes MAY be required to support the capabilities of inserting multiple ELI/ELs and need to advertise the capabilities to the PCEs.

This document proposes the extensions for PCE to perform the computation of the end-to-end path as well as the positions of entropy labels in SR-MPLS networks. The ingress nodes can directly insert the ELI/ELs based on the positions.

4. PCEP Extensions

4.1. The OPEN Object

As defined in [RFC8664], PCEP speakers use SR PCE Capability sub-TLV to exchange information about their SR capability when PST=1 in the PST List of the PATH-SETUP-TYPE-CAPABILITY TLV carried in Open object. This document defines a new flag (E-flag) for SR PCE Capability sub-TLV.

E (ELP Configuration is supported) : A PCE sets this flag bit to 1 carried in Open message to indicate that it supports the computation of SR path with ELP information. A PCC sets this flag to 1 to indicate that it supports the capability of inserting multiple ELI/EL pairs and and supports the results of SR path with ELP from PCE.

4.2. The LSP-EXTENDED-FLAG TLV

The LSP Object is defined in Section 7.3 of [RFC8231]. This document defines a new flag (E-flag) for the LSP-EXTENDED-FLAG TLV carried in LSP Object as defined in [RFC9357].

E (Request for ELP Configuration) : If the bit is set to 1, it indicates that the PCC requests PCE to compute the SR path with ELP information. A PCE would also set this bit to 1 to indicate that the ELP information is included by PCE and encoded in the Path Computation Reply (PCRep) message as per [RFC5440]. And in a stateful PCE model, it also can be carried in Path Computation Update Request (PCUpd) message as per [RFC8231] or LSP Initiate Request (PCInitiate) message as per [RFC8281].

4.3. The SR-ERO Object

SR-ERO subobject is used for SR-TE path which consists of one or more SIDs as defined in [RFC8664]. This document defined a new flag (E-flag) for the SR-ERO subobject.

E (ELP Configuration) : If this flag is set, the PCC SHOULD insert <ELI, EL> into the position after this SR-ERO subobject, otherwise it SHOULD not insert <ELI, EL> after this segment.

5. Operational Example

A PCC can request the computation of SR path and a PCE may respond with PCRep message. And the SR path can also be initiated by PCE with PCInitiate or PCUpd message in stateful PCE mode. When the E bit in LSP object is set to 1 within the message, it indicates to request the ELP configuration with the SR path. The SR path being received by PCC with SR-ERO segment list, for example, <S1, S2, S3, S4, S5, S6>, especially S3 and S6 with E-flag set. It indicates that two <ELI, EL> pairs SHOULD be inserted into the label stack of the SR forwarding entry, respectively after the label for S3 and label for S6. With EL information, the label stack for SR-MPLS would be <label1, label2, label3, ELI, EL, label4, label5, label6, ELI, EL>.

6. Security Considerations

This document defines a new E bit for entropy label, which do not introduce any new security considerations beyond those already listed in [RFC9357], [RFC8662] and [RFC8664].

7. IANA Considerations

7.1. New SR PCE Capability Flag Registry

SR PCE Capability TLV is defined in [RFC8664], and the registry to manage the Flag field of the SR PCE Capability TLV is requested in [RFC8664]. IANA is requested to make allocations from the registry, as follows:

Value	Name	Reference
TBD1	ELP Configuration is supported (E)	[this document]

Table 1

7.2. New LSP-EXTENDED-FLAG Flag Registry

[RFC9357] defines the LSP-EXTENDED-FLAG TLV. IANA is requested to make allocations from the Flag field registry, as follows:

Value	Name	Reference
TBD2	Request for ELP Configuration (E)	[this document]

Table 2

7.3. New SR-ERO Flag Registry

SR-ERO subobject is defined in [RFC8664], and the registry to manage the Flag field of SR-ERO is requested in [RFC8664]. IANA is requested to make allocations from the registry, as follows:

Value	Name	Reference
TBD3	ELP Configuration (E)	[this document]

Table 3

8. Acknowledgements

The authors would like to thank Stephane Litkowski, Dhruv Dhody, Tarek Saad, Zhenbin Li, Jeff Tantsura, Andrew Stone, Xuesong Geng, Ran Chen, Gyan Mishra, Weiqiang Cheng for their review, suggestions and comments to this document.

9. References

9.1. Normative References

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <https://www.rfc-editor.org/info/rfc2119>.

[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>.

[RFC6790] Kompella, K., Drake, J., Amante, S., Henderickx, W., and L. Yong, "The Use of Entropy Labels in MPLS Forwarding", RFC 6790, DOI 10.17487/RFC6790, November 2012, <https://www.rfc-editor.org/info/rfc6790>.

[RFC8174] Leiba, B., "Ambiguity of Uppercase vs Lowercase in RFC 2119 Key Words", BCP 14, RFC 8174, DOI 10.17487/RFC8174, May 2017, <https://www.rfc-editor.org/info/rfc8174>.

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

[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>.

[RFC8476] Tantsura, J., Chunduri, U., Aldrin, S., and P. Psenak, "Signaling Maximum SID Depth (MSD) Using OSPF", RFC 8476, DOI 10.17487/RFC8476, December 2018, <https://www.rfc-editor.org/info/rfc8476>.

[RFC8491] Tantsura, J., Chunduri, U., Aldrin, S., and L. Ginsberg, "Signaling Maximum SID Depth (MSD) Using IS-IS", RFC 8491, DOI 10.17487/RFC8491, November 2018, <https://www.rfc-editor.org/info/rfc8491>.

[RFC8660] Bashandy, A., Ed., Filsfils, C., Ed., Previdi, S., Decraene, B., Litkowski, S., and R. Shakir, "Segment Routing with the MPLS Data Plane", RFC 8660, DOI 10.17487/RFC8660, December 2019, <https://www.rfc-editor.org/info/rfc8660>.

[RFC8662] Kini, S., Kompella, K., Sivabalan, S., Litkowski, S., Shakir, R., and J. Tantsura, "Entropy Label for Source Packet Routing in Networking (SPRING) Tunnels", RFC 8662, DOI 10.17487/RFC8662, December 2019, <https://www.rfc-editor.org/info/rfc8662>.

[RFC8664] Sivabalan, S., Filsfils, C., Tantsura, J., Henderickx, W., and J. Hardwick, "Path Computation Element Communication Protocol (PCEP) Extensions for Segment Routing", RFC 8664, DOI 10.17487/RFC8664, December 2019, <https://www.rfc-editor.org/info/rfc8664>.

[RFC9088] Xu, X., Kini, S., Psenak, P., Filsfils, C., Litkowski, S., and M. Bocci, "Signaling Entropy Label Capability and Entropy Readable Label Depth Using IS-IS", RFC 9088, DOI 10.17487/RFC9088, August 2021, <https://www.rfc-editor.org/info/rfc9088>.

[RFC9089] Xu, X., Kini, S., Psenak, P., Filsfils, C., Litkowski, S., and M. Bocci, "Signaling Entropy Label Capability and Entropy Readable Label Depth Using OSPF", RFC 9089, DOI 10.17487/RFC9089, August 2021, <https://www.rfc-editor.org/info/rfc9089>.

[RFC9357] Xiong, Q., "Label Switched Path (LSP) Object Flag Extension for Stateful PCE", RFC 9357, DOI 10.17487/RFC9357, February 2023, <https://www.rfc-editor.org/info/rfc9357>.

9.2. Informative References

[I-D.ietf-idr-bgp-srmpls-elp]
Liu, Y. and S. Peng, "BGP Extension for SR-MPLS Entropy Label Position", Work in Progress, Internet-Draft, draft-ietf-idr-bgp-srmpls-elp-00, 1 August 2023, <https://datatracker.ietf.org/doc/html/draft-ietf-idr-bgp-srmpls-elp-00>.

Authors' Addresses

Quan Xiong
ZTE Corporation
No.6 Huashi Park Rd
Wuhan
Hubei, 430223
China
Email: xiong.quan@zte.com.cn

Shaofu Peng
ZTE Corporation
No.50 Software Avenue
Nanjing
Jiangsu, 210012
China
Email: peng.shaofu@zte.com.cn

Fengwei Qin
China Mobile
Beijing
China
Email: qinfengwei@chinamobile.com

Junfeng Zhao
CAICT
Beijing
China
Email: zhaojunfeng@caict.ac.cn

Network Working Group
Internet-Draft
Intended status: Informational
Expires: 8 September 2022

C. Schmutzer, Ed.
C. Filsfils
Z. Ali, Ed.
F. Clad
Cisco Systems, Inc.
P. Maheshwari
Airtel India
7 March 2022

Circuit Style Segment Routing Policies
draft-schmutzer-pce-cs-sr-policy-01

Abstract

This document describes how Segment Routing (SR) policies can be used to satisfy the requirements for strict bandwidth guarantees, end-to-end recovery and persistent paths within a segment routing network. SR policies satisfying these requirements are called "circuit-style" SR policies (CS-SR policies).

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 8 September 2022.

Table of Contents

1. Introduction

Segment routing does allow for a single network to carry both typical IP (connection-less) services and connection-oriented transport services. IP services required ECMP and TI-LFA, while transport services that normally are delivered via dedicated circuit-switched SONET/SDH or OTN networks do require:

* Persistent end2end traffic engineered paths that provide predictable and identical latency in both directions

* Strict bandwidth commitment per path to ensure no impact on the Service Level Agreement (SLA) due to changing network load from other services

* End2end protection (<50msec protection switching) and restoration mechanisms

* Monitoring and maintenance of path integrity

* Data plane remaining up while control plane is down

Such a "transport centric" behaviour is referred to as "circuit-style" in this document.

This document describes how SR policies [I-D.ietf-spring-segment-routing-policy] and adjacency-SIDs defined in the SR architecture [RFC8402] together with a stateful Path Computation Element (PCE) [RFC8231] can be used to satisfy those requirements. It includes how end-to-end recovery and path integrity monitoring can be implemented.

SR policies that satisfy those requirements are called "circuit-style" SR policies (CS-SR policies).

2. Terminology

* CS-SR : Circuit-Style Segment Routing

* ID : Identifier

* LSP : Label Switched Path

* LSPA : LSP attributes

* OAM : Operations, Administration and Maintenance

* OF : Objective Function

* PCE : Path Computation Element

* PCEP : Path Computation Element Communication Protocol

* PT : Protection Type

* SID : Segment Identifier

* SLA : Service Level Agreement

* SR : Segment Routing

* STAMP : Simple Two-Way Active Measurement Protocol

* TI-LFA : Topology Independent Loop Free Alternate

* TLV : Type Length Value

3. Reference Model

The reference model for CS-SR policies is following the segment routing architecture [RFC8402] and SR policy architecture [I-D.ietf-spring-segment-routing-policy] and is depicted in Figure 1.

```
                        +--------------+
      +-------------->|     PCE      |<--------------+
      |               +--------------+               |
      |                                              |
      |                                              |
      v   <<<<<<<<<<<<<< CS-SR Policy >>>>>>>>>>>>>  v
+-------+                                       +-------+
|       |=========================================>|       |
|   A   | SR-policy from A to Z                   |   Z   |
|       |<=========================================|       |
+-------+                    SR-policy from Z to A +-------+
```

Figure 1: Circuit-style SR Policy Architecture

By nature of CS-SR policies, paths will be computed and maintained by a stateful PCE defined in [RFC8231]. When using a MPLS data plane [RFC8660], PCEP extensions defined in [RFC8664] will be used. When using a SRv6 data plane [RFC8754], PCEP extensions defined in [I-D.ietf-pce-segment-routing-ipv6] will be used.

In order to satisfy the requirements of CS-SR policies, each link in the topology MUST have:

* An adjacency-SID which is:

  - Manually allocated or persistent : to ensure that its value does not change after a node reload

  - Non-protected : to avoid any local TI-LFA protection to happen upon interface/link failures

* The bandwidth available for CS-SR policies

When using a MPLS data plane [RFC8660] existing IGP extensions defined in [RFC8667] and [RFC8665] and BGP-LS defined in [RFC9085] can be used to distribute the topology information including those persistent and unprotected Adj-SIDs.

When using a SRv6 data plane [RFC8754] the IGP extensions defined in [I-D.ietf-lsr-isis-srv6-extensions] and [I-D.ietf-lsr-ospfv3-srv6-extensions] and BGP-LS extensions in [I-D.ietf-idr-bgpls-srv6-ext] apply.

4. CS-SR Policy Characteristics

A CS-SR policy has the following characteristics:

* Requested bandwidth : bandwidth to be reserved for the CS-SR policy

* Bidirectional co-routed : a CS-SR policy between A and Z is an association of an SR-Policy from A to Z and an SR-Policy from Z to A following the same path(s)

* Deterministic and persistent paths : segment lists with strict hops using unprotected adjacency-SIDs

* Not automatically recomputed or reoptimized : the SID list of a candidate path must not change automatically (for example upon topology change)

* Multiple candidate paths in case of protection/restoration:

  - Following the SR policy architecture, the highest preference valid path is carrying traffic

  - Depending on the protection/restoration scheme (Section 7), lower priority candidate paths

    o may be pre-computed

    o may be pre-programmed

    o may have to be disjoint

* Liveness and performance measurement is activated on each candidate path (Section 6)

5. CS-SR Policy Creation

A CS-SR policy between A and Z is configured both on A (with Z as endpoint) and Z (with A as endpoint) as shown in Figure 1.

Both nodes A and Z act as PCC and delegate path computation to the PCE using the extensions defined in [RFC8664]. The PCRpt message sent from the headends to the PCE contains the following parameters:

* BANDWIDTH object (Section 7.7 of [RFC5440]) : to indicate the requested bandwidth

* LSPA object (section 7.11 of [RFC5440]) : to indicate that no local protection requirements

  - L flag set to 0 : no local protection

  - E flag set to 1 : protection enforcement (section 5 of [I-D.ietf-pce-local-protection-enforcement])

* ASSOCIATION object ([RFC8697]) :

  - Type : Double-sided Bidirectional with Reverse LSP Association ([I-D.ietf-pce-sr-bidir-path])

  - Bidirectional Association Group TLV ([RFC9059]) :

    o R flag is always set to 0 (forward path)

    o C flag is always set to 1 (co-routed)

If the SR-policies are configured with more than one candidate path, a PCEP request is sent per candidate path. Each PCEP request does include the "SR Policy Association" object (type 6) as defined in [I-D.ietf-pce-segment-routing-policy-cp] to make the PCE aware of the candidate path belonging to the same policy.

The signaling extensions described in [I-D.sidor-pce-circuit-style-pcep-extensions] are used to ensure that

* Path determinism is achieved by the PCE only using segment lists representing a strict hop by hop path using unprotected adjacency-SIDs.

* Path persistency across node reloads in the network is achieved by the PCE only including manually configured adj-SIDs in its path computation response.

* Persistency across network changes is achieved by the PCE not performing periodic nor network event triggered re-optimization.

Bandwidth adjustment can be requested after initial creation by signaling both requested and operational bandwidth in the BANDWIDTH object but the PCE is not allowed to respond with a changed path.

6. Operations, Administration, and Maintenance (OAM)

6.1. Liveness

The proper operation of each segment list is validated by both headends using STAMP in loopback measurement mode as described in section 4.2.3 of [I-D.ietf-spring-stamp-srpm].

As the STAMP test packets are including both the segment list of the forward and reverse path, standard segment routing data plane operations will make those packets get switched along the forward path to the tailend and along the reverse path back to the headend.

The headend forms the bidirectional SR Policy association using the procedure described in [I-D.ietf-pce-sr-bidir-path] and receives the information about the reverse segment list from the PCE as described in section 4.5 of [I-D.ietf-pce-multipath]

6.2. Performance Measurement

The same STAMP session used for liveliness monitoring can be used to measure delay. As loopback mode is used only round-trip delay is measured and one-way has to be derived by dividing the round-trip delay by two.

The same STAMP session can also be used to estimate round-trip loss as described in section 5 of [I-D.ietf-spring-stamp-srpm].

7. Recovery Schemes

Various protection and restoration schemes can be implemented. The terms "protection" and "restoration" are used with same subtle distinctions outlined in section 1 of [RFC4872], [RFC4427] and [RFC3386] respectively.

* Protection : another candidate path is computed and fully established in the data plane and ready to carry traffic

* Restoration : a candidate path may be computed and may be partially established but is not ready to carry traffic

7.1. Unprotected

In the most basic scenario no protection nor restoration is required. The CS-SR policy has only one candidate path configured. This candidate path is established, activated (O field in LSP object is set to 2) and is carrying traffic.

In case of a failure the CS-SR policy will go down and traffic will not be recovered.

Typically two CS-SR policies are deployed either within the same network with disjoint paths or in two completely separate networks and the overlay service is responsible for traffic recovery.

7.2. 1+R Restoration

To avoid pre-allocating protection bandwidth in steady state (Section 7.3) but still be able to react to network failures and recover traffic flow in a deterministic way (maintain required bandwidth commitment) the CS-SR policy is configured with two candidate paths.

The candidate path with higher preference is established, activated (O field in LSP object is set to 2) and is carrying traffic.

The second candidate path with lower preference is only established and activated (O field in LSP object is set to 2) upon a failure impacting the first candidate path in order to send traffic over an alternate path through the network around the failure with potentially relaxed constraints but still satisfying the bandwidth commitment.

The second candidate path is generally only requested from the PCE and activated after a failure, but may also be requested and pre-established during CS-SR policy creation with the downside of bandwidth being set aside ahead of time.

As soon as the failure that brought the first candidate path down is cleared, the second candidate path is getting deactivated (O field in LSP object is set to 1) or torn down. The first candidate path is activated (O field in LSP object is set to 2) and traffic sent across it.

Restoration and reversion behavior is bidirectional. As described in Section 6.1, both headends use liveness in loopback mode and therefore even in case of unidirectional failures both headends will detect the failure or clearance of the failure and switch traffic away from the failed or to the recovered candidate path.

7.3. 1:1 Protection

For fast recovery against failures the CS-SR policy is configured with two candidate paths. Both paths are established but only the candidate with higher preference is activated (O field in LSP object is set to 2) and is carrying traffic. The candidate path with lower preference has its O field in LSP object set to 1.

Appropriate routing of the protect path diverse from the working path can be requested from the PCE by using the "Disjointness Association" object (type 2) defined in [RFC8800] in the PCRpt messages. The disjoint requirements are communicated in the "DISJOINTNESS-CONFIGURATION TLV"

* L bit set to 1 for link diversity

* N bit set to 1 for node diversity

* S bit set to 1 for SRLG diversity

* T bit set to enforce strict diversity

The P bit may be set for first candidate path to allow for finding the best working path that does satisfy all constraints without considering diversity to the protect path.

The "Objective Function (OF) TLV" as defined in section 5.3 of [RFC8800] may also be added to minimize the common shared resources.

Upon a failure impacting the candidate path with higher preference carrying traffic, the candidate path with lower preference is activated immediately and traffic is now sent across it.

Protection switching is bidirectional. As described in Section 6.1, both headends will generate and receive their own loopback mode test packets, hence even a unidirectional failure will always be detected by both headends without protection switch coordination required.

Two cases are to be considered when the failure impacting the candidate path with higher preference is cleared:

* Revertive switching : re-activate the candidate path, change O field from 0 to 2 and start sending traffic over it

* Non-revertive switching : do not activate the candidate path, change O field from 0 to 1, keep the second candidate path active with O field set to 2 and continue sending traffic over it

7.4. 1:1+R Protection

For further resiliency in case of multiple concurrent failures that could affect both candidate paths in a Section 7.3 scenario the CS-SR policy is configured with three candidate paths with decreasing preference.

The third candidate path enables restoration and will generally only be established, activated (O field in LSP object is set to 2) and carry traffic after failure(s) have impacted both the candidate path with highest and second highest preference.

The third candidate path may also be requested and pre-computed already whenever either the first or second candidate path went down due to a failure with the downside of bandwidth being set aside ahead of time.

As soon as failure(s) that brought either the first or second candidate path down is cleared the third candidate path is getting deactivated (O field in LSP object is set to 1), the candidate path that recovered is activated (O field in LSP object is set to 2) and traffic sent across it.

Protection switching, restoration and reversion behavior is bidirectional. As described in Section 6.1, both headends use liveness in loopback mode and therefore even in case of unidirectional failures both headends will detect the failure or clearance of the failure and switch traffic away from the failed or to the recovered candidate path.

7.5. External Commands

It is very common to allow operators to trigger a switch between candidate paths even no failure is present. I.e. to proactively drain a resource for maintenance purposes. Operator triggered switching between candidate paths is unidirectional and has to be requested on both headends.

8. Security Considerations

TO BE ADDED

9. IANA Considerations

This document has no IANA actions.

10. Acknowledgements

The author’s want to thank Samuel Sidor, Mike Koldychev, Rakesh Gandhi for providing their review comments.

11. Contributors

Contributors’ Addresses

Brent Foster
Cisco Systems, Inc.
Email: brfoster@cisco.com

Bertrand Duvivier
Cisco System, Inc.
Email: bduvivie@cisco.com

Stephane Litkowski
Cisco Systems, Inc.
Email: slitkows@cisco.com

12. References

12.1. Normative References

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <https://www.rfc-editor.org/info/rfc2119>.

12.2. Informative References

[I-D.ietf-idr-bgpls-srv6-ext]
Dawra, G., Filsfils, C., Talaulikar, K., Chen, M., Bernier, D., and B. Decraene, "BGP Link State Extensions for SRv6", Work in Progress, Internet-Draft, draft-ietf-idr-bgpls-srv6-ext-09, 10 November 2021, <https://www.ietf.org/archive/id/draft-ietf-idr-bgpls-srv6-ext-09.txt>.

[I-D.ietf-lsr-isis-srv6-extensions]
Psenak, P., Filsfils, C., Bashandy, A., Decraene, B., and Z. Hu, "IS-IS Extensions to Support Segment Routing over IPv6 Dataplane", Work in Progress, Internet-Draft, draft-ietf-lsr-isis-srv6-extensions-18, 20 October 2021, <https://www.ietf.org/archive/id/draft-ietf-lsr-isis-srv6-extensions-18.txt>.

[I-D.ietf-lsr-ospfv3-srv6-extensions]
Li, Z., Hu, Z., Cheng, D., Talaulikar, K., and P. Psenak, "OSPFv3 Extensions for SRv6", Work in Progress, Internet-Draft, draft-ietf-lsr-ospfv3-srv6-extensions-03, 19 November 2021, <https://www.ietf.org/archive/id/draft-ietf-lsr-ospfv3-srv6-extensions-03.txt>.

[I-D.ietf-pce-local-protection-enforcement]
Stone, A., Aissaoui, M., Sidor, S., and S. Sivabalan, "Local Protection Enforcement in PCEP", Work in Progress, Internet-Draft, draft-ietf-pce-local-protection-enforcement-04, 30 January 2022, <https://www.ietf.org/archive/id/draft-ietf-pce-local-protection-enforcement-04.txt>.

[I-D.ietf-pce-multipath]
Koldychev, M., Sivabalan, S., Saad, T., Beeram, V. P., Bidgoli, H., Yadav, B., Peng, S., and G. Mishra, "PCEP Extensions for Signaling Multipath Information", Work in Progress, Internet-Draft, draft-ietf-pce-multipath-04, 25 February 2022, <https://www.ietf.org/archive/id/draft-ietf-pce-multipath-04.txt>.

[I-D.ietf-pce-segment-routing-ipv6]
Li, C., Negi, M., Sivabalan, S., Koldychev, M., Kaladharan, P., and Y. Zhu, "PCEP Extensions for Segment Routing leveraging the IPv6 data plane", Work in Progress, Internet-Draft, draft-ietf-pce-segment-routing-ipv6-12, 6 March 2022, <https://www.ietf.org/internet-drafts/draft-ietf-pce-segment-routing-ipv6-12.txt>.

[I-D.ietf-pce-segment-routing-policy-cp]
Koldychev, M., Sivabalan, S., Barth, C., Peng, S., and H. Bidgoli, "PCEP extension to support Segment Routing Policy Candidate Paths", Work in Progress, Internet-Draft, draft-ietf-pce-segment-routing-policy-cp-06, 22 October 2021, <https://www.ietf.org/archive/id/draft-ietf-pce-segment-routing-policy-cp-06.txt>.

[I-D.ietf-pce-sr-bidir-path]
Li, C., Chen, M., Cheng, W., Gandhi, R., and Q. Xiong, "Path Computation Element Communication Protocol (PCEP) Extensions for Associated Bidirectional Segment Routing (SR) Paths", Work in Progress, Internet-Draft, draft-ietf-pce-sr-bidir-path-09, 6 March 2022, <https://www.ietf.org/archive/id/draft-ietf-pce-sr-bidir-path-09.txt>.

[I-D.ietf-spring-segment-routing-policy]
Filsfils, C., Talaulikar, K., Voyer, D., Bogdanov, A., and P. Mattes, "Segment Routing Policy Architecture", Work in Progress, Internet-Draft, draft-ietf-spring-segment-routing-policy-20, 6 March 2022, <https://www.ietf.org/archive/id/draft-ietf-spring-segment-routing-policy-20.txt>.

[I-D.ietf-spring-stamp-srpm]
Gandhi, R., Filsfils, C., Voyer, D., Chen, M., Janssens, B., and R. Foote, "Performance Measurement Using Simple TWAMP (STAMP) for Segment Routing Networks", Work in Progress, Internet-Draft, draft-ietf-spring-stamp-srpm-03, 1 February 2022, <https://www.ietf.org/archive/id/draft-ietf-spring-stamp-srpm-03.txt>.

[I-D.sidor-pce-circuit-style-pcep-extensions]
Sidor, S., Ali, Z., and P. Maheshwari, "PCEP extensions for Circuit Style Policies", Work in Progress, Internet-Draft, draft-sidor-pce-circuit-style-pcep-extensions-00, 7 March 2022, <https://www.ietf.org/archive/id/draft-sidor-pce-circuit-style-pcep-extensions-00.txt>.

[RFC1925] Callon, R., "The Twelve Networking Truths", RFC 1925, DOI 10.17487/RFC1925, April 1996, <https://www.rfc-editor.org/info/rfc1925>.

[RFC3386] Lai, W., Ed. and D. McDysan, Ed., "Network Hierarchy and Multilayer Survivability", RFC 3386, DOI 10.17487/RFC3386, November 2002, <https://www.rfc-editor.org/info/rfc3386>.

[RFC4427] Mannie, E., Ed. and D. Papadimitriou, Ed., "Recovery (Protection and Restoration) Terminology for Generalized Multi-Protocol Label Switching (GMPLS)", RFC 4427, DOI 10.17487/RFC4427, March 2006, <https://www.rfc-editor.org/info/rfc4427>.

[RFC4872] Lang, J.P., Ed., Rekhter, Y., Ed., and D. Papadimitriou, Ed., "RSVP-TE Extensions in Support of End-to-End Generalized Multi-Protocol Label Switching (GMPLS) Recovery", RFC 4872, DOI 10.17487/RFC4872, May 2007, <https://www.rfc-editor.org/info/rfc4872>.

[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>.

[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>.

[RFC8402] Filsfils, C., Ed., Previdi, S., Ed., Ginsberg, L., Decraene, B., Litkowski, S., and R. Shakir, "Segment Routing Architecture", RFC 8402, DOI 10.17487/RFC8402, July 2018, <https://www.rfc-editor.org/info/rfc8402>.

[RFC8660] Bashandy, A., Ed., Filsfils, C., Ed., Previdi, S., Decraene, B., Litkowski, S., and R. Shakir, "Segment Routing with the MPLS Data Plane", RFC 8660, DOI 10.17487/RFC8660, December 2019, <https://www.rfc-editor.org/info/rfc8660>.

[RFC8664] Sivabalan, S., Filsfils, C., Tantsura, J., Henderickx, W., and J. Hardwick, "Path Computation Element Communication Protocol (PCEP) Extensions for Segment Routing", RFC 8664, DOI 10.17487/RFC8664, December 2019, <https://www.rfc-editor.org/info/rfc8664>.

[RFC8665] Psenak, P., Ed., Previdi, S., Ed., Filsfils, C., Gredler, H., Shakir, R., Henderickx, W., and J. Tantsura, "OSPF Extensions for Segment Routing", RFC 8665, DOI 10.17487/RFC8665, December 2019, <https://www.rfc-editor.org/info/rfc8665>.

[RFC8667] Previdi, S., Ed., Ginsberg, L., Ed., Filsfils, C., Bashandy, A., Gredler, H., and B. Decraene, "IS-IS Extensions for Segment Routing", RFC 8667, DOI 10.17487/RFC8667, December 2019, <https://www.rfc-editor.org/info/rfc8667>.

[RFC8697] Minei, I., Crabbe, E., Sivabalan, S., Ananthakrishnan, H., Dhody, D., and Y. Tanaka, "Path Computation Element Communication Protocol (PCEP) Extensions for Establishing Relationships between Sets of Label Switched Paths (LSPs)", RFC 8697, DOI 10.17487/RFC8697, January 2020, <https://www.rfc-editor.org/info/rfc8697>.

[RFC8754] Filsfils, C., Ed., Dukes, D., Ed., Previdi, S., Leddy, J., Matsushima, S., and D. Voyer, "IPv6 Segment Routing Header (SRH)", RFC 8754, DOI 10.17487/RFC8754, March 2020, <https://www.rfc-editor.org/info/rfc8754>.

[RFC8800] Litkowski, S., Sivabalan, S., Barth, C., and M. Negi, "Path Computation Element Communication Protocol (PCEP) Extension for Label Switched Path (LSP) Diversity Constraint Signaling", RFC 8800, DOI 10.17487/RFC8800, July 2020, <https://www.rfc-editor.org/info/rfc8800>.

[RFC9059] Gandhi, R., Ed., Barth, C., and B. Wen, "Path Computation Element Communication Protocol (PCEP) Extensions for Associated Bidirectional Label Switched Paths (LSPs)", RFC 9059, DOI 10.17487/RFC9059, June 2021, <https://www.rfc-editor.org/info/rfc9059>.

[RFC9085] Previdi, S., Talaulikar, K., Ed., Filsfils, C., Gredler, H., and M. Chen, "Border Gateway Protocol - Link State (BGP-LS) Extensions for Segment Routing", RFC 9085, DOI 10.17487/RFC9085, August 2021, <https://www.rfc-editor.org/info/rfc9085>.

Authors' Addresses

Christian Schmutzer (editor)
Cisco Systems, Inc.
Email: cschmutz@cisco.com

Clarence Filsfils
Cisco Systems, Inc.
Email: cfilsfil@cisco.com

Zafar Ali (editor)
Cisco Systems, Inc.
Email: zali@cisco.com

Francois Clad
Cisco Systems, Inc.
Email: fclad@cisco.com

Praveen Maheshwari
Airtel India
Email: Praveen.Maheshwari@airtel.com

Network Working Group
Internet-Draft
Intended status: Informational
Expires: 6 November 2022

C. Schmutzer, Ed.
C. Filsfils
Z. Ali, Ed.
F. Clad
Cisco Systems, Inc.
P. Maheshwari
Airtel India
R. Rokui
Ciena
A. Stone
Nokia
L. Jalil
Verizon
S. Peng
Huawei Technologies
T. Saad
Juniper Networks
D. Voyer
Bell Canada
5 May 2022

Circuit Style Segment Routing Policies
draft-schmutzer-pce-cs-sr-policy-02

Abstract

This document describes how Segment Routing (SR) policies can be used to satisfy the requirements for strict bandwidth guarantees, end-to-end recovery and persistent paths within a segment routing network. SR policies satisfying these requirements are called "circuit-style" SR policies (CS-SR policies).

Status of This Memo

This Internet-Draft will expire on 6 November 2022.

Copyright Notice

Table of Contents

1. Introduction

Segment routing does allow for a single network to carry both typical IP (connection-less) services and connection-oriented transport services commonly referred to as "private lines". IP services typically require ECMP and TI-LFA, while transport services that normally are delivered via dedicated circuit-switched SONET/SDH or OTN networks do require:

* Persistent end-to-end traffic engineered paths that provide predictable and identical latency in both directions

* Strict bandwidth commitment per path to ensure no impact on the Service Level Agreement (SLA) due to changing network load from other services

* End-to-end protection (<50msec protection switching) and restoration mechanisms

* Monitoring and maintenance of path integrity

* Data plane remaining up while control plane is down

Such a "transport centric" behavior is referred to as "circuit-style" in this document.

This document describes how SR policies [I-D.ietf-spring-segment-routing-policy] and the use of adjacency-SIDs defined in the SR architecture [RFC8402] together with a stateful Path Computation Element (PCE) [RFC8231] can be used to satisfy those requirements. It includes how end-to-end recovery and path integrity monitoring can be implemented.

SR policies that satisfy those requirements are called "circuit-style" SR policies (CS-SR policies).

2. Terminology

* BSID : Binding Segment Identifier

* CS-SR : Circuit-Style Segment Routing

* ID : Identifier

* LSP : Label Switched Path

* LSPA : LSP attributes

* OAM : Operations, Administration and Maintenance

* OF : Objective Function

* PCE : Path Computation Element

* PCEP : Path Computation Element Communication Protocol

* PT : Protection Type

* SID : Segment Identifier

* SLA : Service Level Agreement

* SR : Segment Routing

* STAMP : Simple Two-Way Active Measurement Protocol

* TI-LFA : Topology Independent Loop Free Alternate

* TLV : Type Length Value

3. Reference Model

The reference model for CS-SR policies is following the Segment Routing Architecture [RFC8402] and SR Policy Architecture [I-D.ietf-spring-segment-routing-policy] and is depicted in Figure 1.

```
                        +--------------+
      +-------------->|     PCE      |<--------------+
      |               +--------------+               |
      |                                              |
      |                                              |
      v   <<<<<<<<<<<<<< CS-SR Policy >>>>>>>>>>>>>  v
+-------+                                       +-------+
|       |=========================================>|       |
|   A   | SR-policy from A to Z                   |   Z   |
|       |<=========================================|       |
+-------+                   SR-policy from Z to A +-------+
```

Figure 1: Circuit-style SR Policy Reference Model

By nature of CS-SR policies, paths will be computed and maintained by a stateful PCE defined in [RFC8231]. The stateful PCE provides a consistent simple mechanism for initializing the co-routed bidirectional end to end paths, performing bandwidth allocation control, as well as monitoring facilities to ensure SLA compliance for the live of the CS-SR Policy. When using a MPLS data plane

[RFC8660], PCEP extensions defined in [RFC8664] will be used. When using a SRv6 data plane [RFC8754], PCEP extensions defined in [I-D.ietf-pce-segment-routing-ipv6] will be used.

In order to satisfy the requirements of CS-SR policies, each link in the topology MUST have:

* An adjacency-SID which is:

  - Manually allocated or persistent : to ensure that its value does not change after a node reload

  - Non-protected : to avoid any local TI-LFA protection to happen upon interface/link failures

* The bandwidth available for CS-SR policies specified

* A per-hop behavior ([RFC3246] or [RFC2597]) that ensures that the specified bandwidth is available to CS-SR policies at all times independent of any other traffic

When using a MPLS data plane [RFC8660] existing IGP extensions defined in [RFC8667] and [RFC8665] and BGP-LS defined in [RFC9085] can be used to distribute the topology information including those persistent and unprotected adjacency-SIDs.

When using a SRv6 data plane [RFC8754] the IGP extensions defined in [I-D.ietf-lsr-isis-srv6-extensions] and [I-D.ietf-lsr-ospfv3-srv6-extensions] and BGP-LS extensions in [I-D.ietf-idr-bgpls-srv6-ext] apply.

4. CS-SR Policy Characteristics

A CS-SR policy has the following characteristics:

* Requested bandwidth : bandwidth to be reserved for the CS-SR policy

* Bidirectional co-routed : a CS-SR policy between A and Z is an association of an SR-Policy from A to Z and an SR-Policy from Z to A following the same path(s)

* Deterministic and persistent paths : segment lists with strict hops using unprotected adjacency-SIDs

* Not automatically recomputed or reoptimized : the SID list of a candidate path must not change automatically to a SID list representing a different path (for example upon topology change)

* Multiple candidate paths in case of protection/restoration:

  - Following the SR policy architecture, the highest preference valid path is carrying traffic

  - Depending on the protection/restoration scheme (Section 6), lower priority candidate paths

    o may be pre-computed

    o may be pre-programmed

    o may have to be disjoint

* Connectivity verification and performance measurement is activated on each candidate path (Section 7)

5. CS-SR Policy Creation

A CS-SR policy between A and Z is configured both on A (with Z as endpoint) and Z (with A as endpoint) as shown in Figure 1.

Both nodes A and Z act as PCC and delegate path computation to the PCE using the extensions defined in [RFC8664]. The PCRpt message sent from the headends to the PCE contains the following parameters:

* BANDWIDTH object (Section 7.7 of [RFC5440]) : to indicate the requested bandwidth

* LSPA object (section 7.11 of [RFC5440]) : to indicate that no local protection requirements

  - L flag set to 0 : no local protection

  - E flag set to 1 : protection enforcement (section 5 of [I-D.ietf-pce-local-protection-enforcement])

* ASSOCIATION object ([RFC8697]) :

  - Type : Double-sided Bidirectional with Reverse LSP Association ([I-D.ietf-pce-sr-bidir-path])

  - Bidirectional Association Group TLV ([RFC9059]) :

    o R flag is always set to 0 (forward path)

    o C flag is always set to 1 (co-routed)

If the SR-policies are configured with more than one candidate path, a PCEP request is sent per candidate path. Each PCEP request does include the "SR Policy Association" object (type 6) as defined in [I-D.ietf-pce-segment-routing-policy-cp] to make the PCE aware of the candidate path belonging to the same policy.

The signaling extensions described in [I-D.sidor-pce-circuit-style-pcep-extensions] are used to ensure that

* Path determinism is achieved by the PCE only using segment lists representing a strict hop by hop path using unprotected adjacency-SIDs.

* Path persistency across node reloads in the network is achieved by the PCE only including manually configured adjacency-SIDs in its path computation response.

* Persistency across network changes is achieved by the PCE not performing periodic nor network event triggered re-optimization.

Bandwidth adjustment can be requested after initial creation by signaling both requested and operational bandwidth in the BANDWIDTH object but the PCE is not allowed to respond with a changed path.

As discussed in section 3.2 of [I-D.ietf-pce-multipath] it may be necessary to use load-balancing across multiple paths to satisfy the bandwidth requirement of a candidate path. In such a case the PCE will notify the PCC to install multiple segment lists using the signaling procedures described in section 5.3 of [I-D.ietf-pce-multipath].

5.1. Maximum Segment Depth

A Segment Routed path defined by a segment list is constrained by maximum segment depth (MSD), which is the maximum number of segments a router can impose onto a packet. [RFC8491], [RFC8476], [RFC8814] and [RFC8664] provide the necessary capabilities for a PCE to determine the MSD capability of a router. The MSD constraint is typically resolved by leveraging a label stack reduction technique, such as using Node SIDs and/or BSIDs (SR architecture [RFC8402]) in a segment list, which represents one or many hops in a given path.

As described in Section 4, adjacency-SIDs without local protection are to be used for CS-SR policies to ensure no ECMP, no rerouting due to topological changes nor localized protection is being invoked on the traffic, as the alternate path may not be providing the desired SLA.

If a CS-SR Policy path requires SID List reduction, a Node SID cannot be utilized as it is eligible for traffic rerouting following IGP re-convergence. However, a BSID can be programmed to a transit node, if the following requirements are met:

* The BSID is unprotected, hence only has one candidate path

* The BSID follows the rerouting and optimization characteristics defined in Section 4 which implies the SID list of the candidate path MUST only use unprotected adjacency-SIDs.

This ensures that any CS-SR policies in which the BSID provides transit for do not get rerouted due to topological changes or protected due to failures. A BSID may be pre-programmed in the network or automatically injected in the network by a PCE.

6. Recovery Schemes

Various protection and restoration schemes can be implemented. The terms "protection" and "restoration" are used with the same subtle distinctions outlined in section 1 of [RFC4872], [RFC4427] and [RFC3386] respectively.

* Protection : another candidate path is computed and fully established in the data plane and ready to carry traffic

* Restoration : a candidate path may be computed and may be partially established but is not ready to carry traffic

The term "failure" is used to represent both "hard failures" such complete loss of connectivity detected by Section 7.1 or degradation, a packet loss ratio, beyond a configured acceptable threshold.

6.1. Unprotected

In the most basic scenario no protection nor restoration is required. The CS-SR policy has only one candidate path configured. This candidate path is established, activated (O field in LSP object is set to 2) and is carrying traffic.

In case of a failure the CS-SR policy will go down and traffic will not be recovered.

Typically two CS-SR policies are deployed either within the same network with disjoint paths or in two completely separate networks and the overlay service is responsible for traffic recovery.

6.2. 1:1 Protection

For fast recovery against failures the CS-SR policy is configured with two candidate paths. Both paths are established but only the candidate with higher preference is activated (O field in LSP object is set to 2) and is carrying traffic. The candidate path with lower preference has its O field in LSP object set to 1.

Appropriate routing of the protect path diverse from the working path can be requested from the PCE by using the "Disjointness Association" object (type 2) defined in [RFC8800] in the PCRpt messages. The disjoint requirements are communicated in the "DISJOINTNESS-CONFIGURATION TLV"

* L bit set to 1 for link diversity

* N bit set to 1 for node diversity

* S bit set to 1 for SRLG diversity

* T bit set to enforce strict diversity

The P bit may be set for first candidate path to allow for finding the best working path that does satisfy all constraints without considering diversity to the protect path.

The "Objective Function (OF) TLV" as defined in section 5.3 of [RFC8800] may also be added to minimize the common shared resources.

Upon a failure impacting the candidate path with higher preference carrying traffic, the candidate path with lower preference is activated immediately and traffic is now sent across it.

Protection switching is bidirectional. As described in Section 7.1, both headends will generate and receive their own loopback mode test packets, hence even a unidirectional failure will always be detected by both headends without protection switch coordination required.

Two cases are to be considered when the failure impacting the candidate path with higher preference is cleared:

* Revertive switching : re-activate the candidate path, change O field from 0 to 2 and start sending traffic over it

* Non-revertive switching : do not activate the candidate path, change O field from 0 to 1, keep the second candidate path active with O field set to 2 and continue sending traffic over it

6.3. Restoration

6.3.1. 1+R Restoration

Compared to 1:1 protection described in Section 6.2, this restoration scheme avoids pre-allocating protection bandwidth in steady state, while still being able to recover traffic flow in case of a network failure in a deterministic way (maintain required bandwidth commitment)

The CS-SR policy is configured with two candidate paths. The candidate path with higher preference is established, activated (O field in LSP object is set to 2) and is carrying traffic.

The second candidate path with lower preference is only established and activated (O field in LSP object is set to 2) upon a failure impacting the first candidate path in order to send traffic over an alternate path through the network around the failure with potentially relaxed constraints but still satisfying the bandwidth commitment.

The second candidate path is generally only requested from the PCE and activated after a failure, but may also be requested and pre-established during CS-SR policy creation with the downside of bandwidth being set aside ahead of time.

As soon as failure(s) that brought the first candidate path down are cleared, the second candidate path is getting deactivated (O field in LSP object is set to 1) or torn down. The first candidate path is activated (O field in LSP object is set to 2) and traffic sent across it.

Restoration and reversion behavior is bidirectional. As described in Section 7.1, both headends use connectivity verification in loopback mode and therefore even in case of unidirectional failures both headends will detect the failure or clearance of the failure and switch traffic away from the failed or to the recovered candidate path.

6.3.2. 1:1+R Restoration

For further resiliency in case of multiple concurrent failures that could affect both candidate paths of 1:1 protection described in Section 6.2, a third candidate path with a preference lower than the other two candidate paths is added to the CS-SR policy.

The third candidate path enables restoration and will generally only be established, activated (O field in LSP object is set to 2) and carry traffic after failure(s) have impacted both the candidate path with highest and second highest preference.

The third candidate path may also be requested and pre-computed already whenever either the first or second candidate path went down due to a failure with the downside of bandwidth being set aside ahead of time.

As soon as failure(s) that brought either the first or second candidate path down are cleared the third candidate path is getting deactivated (O field in LSP object is set to 1), the candidate path that recovered is activated (O field in LSP object is set to 2) and traffic sent across it.

Again restoration and reversion behavior is bidirectional. As described in Section 7.1, both headends use connectivity verification in loopback mode and therefore even in case of unidirectional failures both headends will detect the failure or clearance of the failure and switch traffic away from the failed or to the recovered candidate path.

7. Operations, Administration, and Maintenance (OAM)

7.1. Connectivity Verification

The proper operation of each segment list is validated by both headends using STAMP in loopback measurement mode as described in section 4.2.3 of [I-D.ietf-spring-stamp-srpm].

As the STAMP test packets are including both the segment list of the forward and reverse path, standard segment routing data plane operations will make those packets get switched along the forward path to the tailend and along the reverse path back to the headend.

The headend forms the bidirectional SR Policy association using the procedure described in [I-D.ietf-pce-sr-bidir-path] and receives the information about the reverse segment list from the PCE as described in section 4.5 of [I-D.ietf-pce-multipath]

7.2. Performance Measurement

The same STAMP session is used to estimate round-trip loss as described in section 5 of [I-D.ietf-spring-stamp-srpm].

The same STAMP session used for connectivity verification can be used to measure delay. As loopback mode is used only round-trip delay is measured and one-way has to be derived by dividing the round-trip delay by two.

7.3. Candidate Path Validity Verification

A stateful PCE is in sync with the network topology and the CS-SR Policies provisioned on the headend routers. As described in Section 4 a path must not be automatically recomputed after or optimized for topology changes. However there may be a requirement for a PCE to tear down a path if the path no longer satisfies the original requirements, detected by PCE, such as insufficient bandwidth, diversity constraint no longer met or latency constraint exceeded.

The PCC may measure the actual bandwidth utilization of a CS-SR policy and report it to the PCE in order for the PCE to take an appropriate action if necessary.

For a CS-SR policy configured with multiple candidate paths, a PCC may switch to another candidate path if the PCE decided to tear down the active candidate path.

8. External Commands

8.1. Candidate Path Switchover

It is very common to allow operators to trigger a switch between candidate paths even if no failure is present. I.e. to proactively drain a resource for maintenance purposes. Operator triggered switching between candidate paths is unidirectional and has to be requested on both headends.

8.2. Candidate Path Recomputation

While no automatic re-optimization or pre-computation of CS-SR policy candidate paths is allowed as specified in Section 4, network operators trying to optimize network utilization may explicitly request a candidate path to be re-computed at a certain point in time.

9. Security Considerations

TO BE ADDED

10. IANA Considerations

This document has no IANA actions.

11. Acknowledgements

The author's want to thank Samuel Sidor, Mike Koldychev, Rakesh Gandhi and Tarek Saad for providing their review comments.

12. Contributors

Contributors' Addresses

Brent Foster
Cisco Systems, Inc.
Email: brfoster@cisco.com

Bertrand Duvivier
Cisco System, Inc.
Email: bduvivie@cisco.com

Stephane Litkowski
Cisco Systems, Inc.
Email: slitkows@cisco.com

Jie Dong
Huawei Technologies
Email: jie.dong@huawei.com

13. References

13.1. Normative References

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <https://www.rfc-editor.org/info/rfc2119>.

13.2. Informative References

[I-D.ietf-idr-bgpls-srv6-ext] Dawra, G., Filsfils, C., Talaulikar, K., Chen, M., Bernier, D., and B. Decraene, "BGP Link State Extensions for SRv6", Work in Progress, Internet-Draft, draft-ietf-idr-bgpls-srv6-ext-09, 10 November 2021, <https://www.ietf.org/archive/id/draft-ietf-idr-bgpls-srv6-ext-09.txt>.

[I-D.ietf-lsr-isis-srv6-extensions]
Psenak, P., Filsfils, C., Bashandy, A., Decraene, B., and Z. Hu, "IS-IS Extensions to Support Segment Routing over IPv6 Dataplane", Work in Progress, Internet-Draft, draft-ietf-lsr-isis-srv6-extensions-18, 20 October 2021, <https://www.ietf.org/archive/id/draft-ietf-lsr-isis-srv6-extensions-18.txt>.

[I-D.ietf-lsr-ospfv3-srv6-extensions]
Li, Z., Hu, Z., Cheng, D., Talaulikar, K., and P. Psenak, "OSPFv3 Extensions for SRv6", Work in Progress, Internet-Draft, draft-ietf-lsr-ospfv3-srv6-extensions-03, 19 November 2021, <https://www.ietf.org/archive/id/draft-ietf-lsr-ospfv3-srv6-extensions-03.txt>.

[I-D.ietf-pce-local-protection-enforcement]
Stone, A., Aissaoui, M., Sidor, S., and S. Sivabalan, "Local Protection Enforcement in PCEP", Work in Progress, Internet-Draft, draft-ietf-pce-local-protection-enforcement-05, 4 May 2022, <https://www.ietf.org/archive/id/draft-ietf-pce-local-protection-enforcement-05.txt>.

[I-D.ietf-pce-multipath]
Koldychev, M., Sivabalan, S., Saad, T., Beeram, V. P., Bidgoli, H., Yadav, B., Peng, S., and G. Mishra, "PCEP Extensions for Signaling Multipath Information", Work in Progress, Internet-Draft, draft-ietf-pce-multipath-05, 30 March 2022, <https://www.ietf.org/archive/id/draft-ietf-pce-multipath-05.txt>.

[I-D.ietf-pce-segment-routing-ipv6]
Li, C., Negi, M., Sivabalan, S., Koldychev, M., Kaladharan, P., and Y. Zhu, "PCEP Extensions for Segment Routing leveraging the IPv6 data plane", Work in Progress, Internet-Draft, draft-ietf-pce-segment-routing-ipv6-13, 1 April 2022, <https://www.ietf.org/internet-drafts/draft-ietf-pce-segment-routing-ipv6-13.txt>.

[I-D.ietf-pce-segment-routing-policy-cp]
Koldychev, M., Sivabalan, S., Barth, C., Peng, S., and H. Bidgoli, "PCEP extension to support Segment Routing Policy Candidate Paths", Work in Progress, Internet-Draft, draft-ietf-pce-segment-routing-policy-cp-07, 21 April 2022, <https://www.ietf.org/archive/id/draft-ietf-pce-segment-routing-policy-cp-07.txt>.

[I-D.ietf-pce-sr-bidir-path]
Li, C., Chen, M., Cheng, W., Gandhi, R., and Q. Xiong, "Path Computation Element Communication Protocol (PCEP) Extensions for Associated Bidirectional Segment Routing (SR) Paths", Work in Progress, Internet-Draft, draft-ietf-pce-sr-bidir-path-09, 6 March 2022, <https://www.ietf.org/archive/id/draft-ietf-pce-sr-bidir-path-09.txt>.

[I-D.ietf-spring-segment-routing-policy]
Filsfils, C., Talaulikar, K., Voyer, D., Bogdanov, A., and P. Mattes, "Segment Routing Policy Architecture", Work in Progress, Internet-Draft, draft-ietf-spring-segment-routing-policy-22, 22 March 2022, <https://www.ietf.org/archive/id/draft-ietf-spring-segment-routing-policy-22.txt>.

[I-D.ietf-spring-stamp-srpm]
Gandhi, R., Filsfils, C., Voyer, D., Chen, M., Janssens, B., and R. Foote, "Performance Measurement Using Simple TWAMP (STAMP) for Segment Routing Networks", Work in Progress, Internet-Draft, draft-ietf-spring-stamp-srpm-03, 1 February 2022, <https://www.ietf.org/archive/id/draft-ietf-spring-stamp-srpm-03.txt>.

[I-D.sidor-pce-circuit-style-pcep-extensions]
Sidor, S., Ali, Z., and P. Maheshwari, "PCEP extensions for Circuit Style Policies", Work in Progress, Internet-Draft, draft-sidor-pce-circuit-style-pcep-extensions-00, 7 March 2022, <https://www.ietf.org/archive/id/draft-sidor-pce-circuit-style-pcep-extensions-00.txt>.

[RFC1925] Callon, R., "The Twelve Networking Truths", RFC 1925, DOI 10.17487/RFC1925, April 1996, <https://www.rfc-editor.org/info/rfc1925>.

[RFC2597] Heinanen, J., Baker, F., Weiss, W., and J. Wroclawski, "Assured Forwarding PHB Group", RFC 2597, DOI 10.17487/RFC2597, June 1999, <https://www.rfc-editor.org/info/rfc2597>.

[RFC3246] Davie, B., Charny, A., Bennet, J.C.R., Benson, K., Le Boudec, J.Y., Courtney, W., Davari, S., Firoiu, V., and D. Stiliadis, "An Expedited Forwarding PHB (Per-Hop Behavior)", RFC 3246, DOI 10.17487/RFC3246, March 2002, <https://www.rfc-editor.org/info/rfc3246>.

[RFC3386] Lai, W., Ed. and D. McDysan, Ed., "Network Hierarchy and Multilayer Survivability", RFC 3386, DOI 10.17487/RFC3386, November 2002, <https://www.rfc-editor.org/info/rfc3386>.

[RFC4427] Mannie, E., Ed. and D. Papadimitriou, Ed., "Recovery (Protection and Restoration) Terminology for Generalized Multi-Protocol Label Switching (GMPLS)", RFC 4427, DOI 10.17487/RFC4427, March 2006, <https://www.rfc-editor.org/info/rfc4427>.

[RFC4872] Lang, J.P., Ed., Rekhter, Y., Ed., and D. Papadimitriou, Ed., "RSVP-TE Extensions in Support of End-to-End Generalized Multi-Protocol Label Switching (GMPLS) Recovery", RFC 4872, DOI 10.17487/RFC4872, May 2007, <https://www.rfc-editor.org/info/rfc4872>.

[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>.

[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>.

[RFC8402] Filsfils, C., Ed., Previdi, S., Ed., Ginsberg, L., Decraene, B., Litkowski, S., and R. Shakir, "Segment Routing Architecture", RFC 8402, DOI 10.17487/RFC8402, July 2018, <https://www.rfc-editor.org/info/rfc8402>.

[RFC8476] Tantsura, J., Chunduri, U., Aldrin, S., and P. Psenak, "Signaling Maximum SID Depth (MSD) Using OSPF", RFC 8476, DOI 10.17487/RFC8476, December 2018, <https://www.rfc-editor.org/info/rfc8476>.

[RFC8491] Tantsura, J., Chunduri, U., Aldrin, S., and L. Ginsberg, "Signaling Maximum SID Depth (MSD) Using IS-IS", RFC 8491, DOI 10.17487/RFC8491, November 2018, <https://www.rfc-editor.org/info/rfc8491>.

[RFC8660] Bashandy, A., Ed., Filsfils, C., Ed., Previdi, S., Decraene, B., Litkowski, S., and R. Shakir, "Segment Routing with the MPLS Data Plane", RFC 8660, DOI 10.17487/RFC8660, December 2019, <https://www.rfc-editor.org/info/rfc8660>.

[RFC8664] Sivabalan, S., Filsfils, C., Tantsura, J., Henderickx, W., and J. Hardwick, "Path Computation Element Communication Protocol (PCEP) Extensions for Segment Routing", RFC 8664, DOI 10.17487/RFC8664, December 2019, <https://www.rfc-editor.org/info/rfc8664>.

[RFC8665] Psenak, P., Ed., Previdi, S., Ed., Filsfils, C., Gredler, H., Shakir, R., Henderickx, W., and J. Tantsura, "OSPF Extensions for Segment Routing", RFC 8665, DOI 10.17487/RFC8665, December 2019, <https://www.rfc-editor.org/info/rfc8665>.

[RFC8667] Previdi, S., Ed., Ginsberg, L., Ed., Filsfils, C., Bashandy, A., Gredler, H., and B. Decraene, "IS-IS Extensions for Segment Routing", RFC 8667, DOI 10.17487/RFC8667, December 2019, <https://www.rfc-editor.org/info/rfc8667>.

[RFC8697] Minei, I., Crabbe, E., Sivabalan, S., Ananthakrishnan, H., Dhody, D., and Y. Tanaka, "Path Computation Element Communication Protocol (PCEP) Extensions for Establishing Relationships between Sets of Label Switched Paths (LSPs)", RFC 8697, DOI 10.17487/RFC8697, January 2020, <https://www.rfc-editor.org/info/rfc8697>.

[RFC8754] Filsfils, C., Ed., Dukes, D., Ed., Previdi, S., Leddy, J., Matsushima, S., and D. Voyer, "IPv6 Segment Routing Header (SRH)", RFC 8754, DOI 10.17487/RFC8754, March 2020, <https://www.rfc-editor.org/info/rfc8754>.

[RFC8800] Litkowski, S., Sivabalan, S., Barth, C., and M. Negi, "Path Computation Element Communication Protocol (PCEP) Extension for Label Switched Path (LSP) Diversity Constraint Signaling", RFC 8800, DOI 10.17487/RFC8800, July 2020, <https://www.rfc-editor.org/info/rfc8800>.

[RFC8814] Tantsura, J., Chunduri, U., Talaulikar, K., Mirsky, G., and N. Triantafillis, "Signaling Maximum SID Depth (MSD) Using the Border Gateway Protocol - Link State", RFC 8814, DOI 10.17487/RFC8814, August 2020, <https://www.rfc-editor.org/info/rfc8814>.

[RFC9059] Gandhi, R., Ed., Barth, C., and B. Wen, "Path Computation Element Communication Protocol (PCEP) Extensions for Associated Bidirectional Label Switched Paths (LSPs)", RFC 9059, DOI 10.17487/RFC9059, June 2021, <https://www.rfc-editor.org/info/rfc9059>.

[RFC9085] Previdi, S., Talaulikar, K., Ed., Filsfils, C., Gredler, H., and M. Chen, "Border Gateway Protocol - Link State (BGP-LS) Extensions for Segment Routing", RFC 9085, DOI 10.17487/RFC9085, August 2021, <https://www.rfc-editor.org/info/rfc9085>.

Authors' Addresses

Christian Schmutzer (editor)
Cisco Systems, Inc.
Email: cschmutz@cisco.com

Clarence Filsfils
Cisco Systems, Inc.
Email: cfilsfil@cisco.com

Zafar Ali (editor)
Cisco Systems, Inc.
Email: zali@cisco.com

Francois Clad
Cisco Systems, Inc.
Email: fclad@cisco.com

Praveen Maheshwari
Airtel India
Email: Praveen.Maheshwari@airtel.com

Reza Rokui
Ciena
Email: rrokui@ciena.com

Andrew Stone
Nokia
Email: andrew.stone@nokia.com

Luay Jalil
Verizon
Email: luay.jalil@verizon.com

Shuping Peng
Huawei Technologies
Email: pengshuping@huawei.com

Tarek Saad
Juniper Networks
Email: tsaad@juniper.net

Daniel Voyer
Bell Canada
Email: daniel.voyer@bell.ca

PCE Working Group
Internet-Draft
Intended status: Standards Track
Expires: 5 September 2022

S. Sidor
Z. Ali
Cisco Systems, Inc.
P. Maheshwari
Airtel India
4 March 2022

PCEP extensions for Circuit Style Policies
draft-sidor-pce-circuit-style-pcep-extensions-00

Abstract

This document proposes a set of extensions for Path Computation Element Communication Protocol (PCEP) for Circuit Style Policies - Segment-Routing Policy designed to satisfy requirements for connection-oriented transport services. New TLV is introduced to control path recomputation triggers and new flag to add ability to request path with strict hops only.

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

Table of Contents

1. Introduction

Usage of Segment-routing and PCEP in connection-oriented transport services require path persistancy and hop-by-hop behavior for PCE computed paths.

Circuit-Style Policy introduced in [I-D.schmutzer-pce-cs-sr-policy] requires PCEP extensions, which are covered in this document.

This document:

* Introduces possibility to request strict path from the PCE by extending LSP-EXTENDED-FLAG TLV

* Adding new TLV for encoding blocked path recomputation triggers to the PCE is introduced, to be carried inside the LSP object, which is defined in [RFC8231].

* Clarifies usage of existing O-flag from RP object in Segment-routing

PCEP extensions described in this document are applicable to RSVP-TE and SR-TE.

2. Terminology

The following terminologies are used in this document:

ERO: Explicit Route Object

IGP: Interior Gateway Protocol

LSP: Label Switched Path.

LSPA: Label Switched Path Attributes.

OTN: Optical Transport Network.

PCC: Path Computation Client

PCE: Path Computation Element

PCEP: Path Computation Element Protocol.

SDH: Synchronous Digital Hierarchy

SID: Segment Identifier

SONET: Synchronous Optical Network

SR: Segment Routing.

SR-TE: Segment Routing Traffic Engineering.

3. Overview of Extensions to PCEP

3.1. LSP-EXTENDED-FLAG TLV

O-flag is proposed in the LSP-EXTENDED-FLAG TLV, which was introduced in 5.1.2 of [I-D.ietf-pce-lsp-extended-flags] and extended with E-flag in [I-D.peng-pce-entropy-label-position]

The format of the LSP-EXTENDED-FLAG 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
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|           Type=TBD1           |          Length               |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                                                             |O|E|
//                 LSP Extended Flags                          //
|                                                               |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 1: LSP-EXTENDED-FLAG TLV Format

Type (16 bits): the value is TBD1 by IANA.

Length (16 bits): multiple of 4 octets.

O (Strict-Path): If set to 1, this indicates to the PCE that a path exclusively made of strict hops is required. Strict hop definition can be found in Section 4.1

3.2. RECOMPUTATION-TRIGGERS TLV

This document defines new TLV for the LSP Object for encoding information about blocked path recomputation triggers.

```
 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 = TBD2         |             Length = 4        |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|             Reserved          |      Flags                  |T|P|
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Type (16 bits): the value is TBD2 by IANA.

Length (16 bits): 4 octets

Reserved: MUST be set to zero by the sender and MUST be ignored by the receiver.

Flags: This document defines the following flag bits. The other bits MUST be set to zero by the sender and MUST be ignored by the receiver.

* T (Topology-change): If set to 1, the PCE MUST NOT trigger recomputation as a result of received updated topology information.

* P (Periodic-timer): If set to 1, the PCE MUST NOT trigger recomputation based on any periodic timer.

4. Operation

4.1. Strict path enforcement

PCC MAY set the O flag in LSP-EXTENDED-FLAG TLV in PCRpt message to the PCE to indicate that a path exclusively made of strict hops is required.

O flag cleared or LSP-EXTENDED-FLAG TLV not included indicates that a loose path is acceptable.

In PCUpdate or PCInitiate messages, when the O bit is set, this indicates that strict path is provided.

The flag is applicable only for stateful messages. Existing O flag in RP object MAY be used to indicate similar behavior in PCReq and PCRep messages as described in as described in Section 7.4.1 of [RFC5440].

If O flag is set to 1 for both stateful and stateless messages for SR paths introduced in [RFC8664], the PCE MUST use Adjacency SIDs only.

4.2. Path computation triggers

PCC MAY set flags in RECOMPUTATION-TRIGGERS-TLV to block specific triggers. If TLV is not included or all flags are set to 0, then the PCE MAY use any event to start path computation.

Disabled recomputation triggered by topology event is not blocking path computation started based PCRpt or based on updated state of associated LSP.

If trigger blocked by specific flag is not supported or allowed on the PCE, then PCE MAY ignore received flag value. The PCE SHOULD reflect blocked triggers in PCUpdate message.

TLV MAY be included in PCInitiate and PCUpdate messages to indicate, which triggers will be disabled on the PCE. PCC should reflect flag values in PCRpt messages to forward requirement to other PCEs in the network.

5. Security Considerations

No additional security measure is required.

6. IANA Considerations

6.1. LSP-EXTENDED-FLAG TLV

[I-D.ietf-pce-lsp-extended-flags] defines the LSP-EXTENDED-FLAG TLV. IANA is requested to make the following assignment from the "LSP-EXTENDED-FLAG TLV Flag Field" registry:

Bit	Description	Reference
TBD1	Strict-Path Flag (O)	This document

Table 1

6.2. RECOMPUTATION-TRIGGERS TLV

IANA is requested to make the assignment of a new value for the existing "PCEP TLV Type Indicators" registry as follows:

TLV Type	TLV Name	Reference
TBD2	RECOMPUTATION-TRIGGERS TLV	This document

Table 2

7. References

7.1. Normative References

[I-D.ietf-pce-lsp-extended-flags]
Xiong, Q., "LSP Object Flag Extension of Stateful PCE", Work in Progress, Internet-Draft, draft-ietf-pce-lsp-extended-flags-01, 18 October 2021, <https://www.ietf.org/archive/id/draft-ietf-pce-lsp-extended-flags-01.txt>.

[I-D.peng-pce-entropy-label-position]
Xiong, Q., Peng, S., and F. Qin, "PCEP Extension for SR-MPLS Entropy Label Position", Work in Progress, Internet-Draft, draft-peng-pce-entropy-label-position-07, 2 March 2022, <https://www.ietf.org/archive/id/draft-peng-pce-entropy-label-position-07.txt>.

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <https://www.rfc-editor.org/info/rfc2119>.

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

[RFC8174] Leiba, B., "Ambiguity of Uppercase vs Lowercase in RFC 2119 Key Words", BCP 14, RFC 8174, DOI 10.17487/RFC8174, May 2017, <https://www.rfc-editor.org/info/rfc8174>.

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

[RFC8664] Sivabalan, S., Filsfils, C., Tantsura, J., Henderickx, W., and J. Hardwick, "Path Computation Element Communication Protocol (PCEP) Extensions for Segment Routing", RFC 8664, DOI 10.17487/RFC8664, December 2019, <https://www.rfc-editor.org/info/rfc8664>.

7.2. Informative References

[I-D.schmutzer-pce-cs-sr-policy]
Schmutzer, C., Filsfils, C., Ali, Z., and F. Clad, "Circuit Style Segment Routing Policies", Work in Progress, Internet-Draft, draft-schmutzer-pce-cs-sr-policy-00, 30 September 2021, <https://www.ietf.org/archive/id/draft-schmutzer-pce-cs-sr-policy-00.txt>.

Authors' Addresses

Samuel Sidor
Cisco Systems, Inc.
Eurovea Central 3.

Pribinova 10
811 09 Bratislava
Slovakia
Email: ssidor@cisco.com

Zafar Ali
Cisco Systems, Inc.
Email: zali@cisco.com

Praveen Maheshwari
Airtel India
Email: Praveen.Maheshwari@airtel.com

PCE Working Group
Internet-Draft
Intended status: Standards Track
Expires: 17 June 2024

S. Sidor
Cisco Systems, Inc.
P. Maheshwari
Airtel India
A. Stone
Nokia
L. Jalil
Verizon
S. Peng
Huawei Technologies
15 December 2023

PCEP extensions for Circuit Style Policies
draft-sidor-pce-circuit-style-pcep-extensions-06

Abstract

This document proposes a set of extensions for Path Computation Element Communication Protocol (PCEP) for Circuit Style Policies - Segment-Routing Policy designed to satisfy requirements for connection-oriented transport services. New TLV is introduced to control path recomputation and new flag to add ability to request path with strict hops only.

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 will expire on 17 June 2024.

Copyright Notice

Table of Contents

1. Introduction

Usage of Segment-routing and PCEP in connection-oriented transport services require path persistancy and hop-by-hop behavior for PCE computed paths.

Circuit-Style Policy introduced in [I-D.ietf-spring-cs-sr-policy] requires PCEP extensions, which are covered in this document.

This document:

* Introduces possibility to request strict path from the PCE by extending LSP-EXTENDED-FLAG TLV

* Adding new TLV to encode information about disabling path recomputation for specific path to the PCE, to be carried inside the LSPA object, which is defined in [RFC5440].

* Clarifies usage of existing O-flag from RP object in Segment-routing

PCEP extensions described in this document can be used with any Path Setup Type.

2. Terminology

The following terminologies are used in this document:

ERO: Explicit Route Object

IGP: Interior Gateway Protocol

LSP: Label Switched Path.

LSPA: Label Switched Path Attributes.

OTN: Optical Transport Network.

PCC: Path Computation Client

PCE: Path Computation Element

PCEP: Path Computation Element Protocol.

SDH: Synchronous Digital Hierarchy

SID: Segment Identifier

SONET: Synchronous Optical Network

SR: Segment Routing.

SR-TE: Segment Routing Traffic Engineering.

3. Overview of Extensions to PCEP

3.1. New flag in the LSP-EXTENDED-FLAG TLV

O-flag is proposed in the LSP-EXTENDED-FLAG TLV, which was introduced in 3.1 of [RFC9357]. TLV format will be added after assigning O-flag bit position by IANA.

O (Strict-Path): If set to 1, this indicates to the PCE that a path exclusively made of strict hops is required. Strict hop definition is described in Section 4.1

3.2. PATH-RECOMPUTATION TLV

This document defines new TLV for the LSPA Object for encoding information whether path recomputation is allowed for delegated LSP. The TLV is optional. If the TLV is included in LSPA object, the PCE MUST NOT recompute path in cases specified by flags in the TLV.

```
 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 = TBD2         |             Length = 4        |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|             Reserved          |      Flags                  |P|F|
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Type (16 bits): the value is TBD2 by IANA.

Length (16 bits): 4 octets

Reserved: MUST be set to zero by the sender and MUST be ignored by the receiver.

Flags: This document defines the following flag bits. The other bits MUST be set to zero by the sender and MUST be ignored by the receiver.

* P (Permanent): If set to 1, the PCE MUST NOT recompute path even if current path is not satisfying path computation constraints. If this flag is cleared, then the PCE SHOULD recompute path if original path is invalidated.

* F (Force): If set to 1, the PCE MUST NOT update path. If flag is cleared, the PCE MAY update path based on explicit request from operator.

4. Operation

4.1. Strict path enforcement

PCC MAY set the O flag in LSP-EXTENDED-FLAG TLV in PCRpt message to the PCE to indicate that a path exclusively made of strict hops is required.

O flag cleared or LSP-EXTENDED-FLAG TLV not included indicates that a loose path is acceptable.

In PCUpdate or PCInitiate messages, PCE MAY set O bit if strict path is provided.

The flag is applicable only for stateful messages. Existing O flag in RP object MAY be used to indicate similar behavior in PCReq and PCRep messages as described in as described in Section 7.4.1 of [RFC5440].

If O flag is set to 1 for both stateful and stateless messages for SR paths introduced in [RFC8664], PCE MUST use only SIDs, which will use explicitly specified adjacencies for packet forwarding. For example Adjacency SIDs MAY be used, but Prefix SIDs MUST NOT be used (even if there is only one adjacency).

4.2. Path recomputation

PCC MAY set flags in PATH-RECOMPUTATION TLV to control path computation behavior on PCE side. If TLV is not included, then the PCE MAY use local policy to trigger path-computation or LSP path update.

The presence of the TLV is blocking path recomputation based on various triggers like topology update, any periodic update or changed state of other LSPs in the network. LSP path MAY be modified if forwarded packets will still use same path - for example if same path can be encoded using Adjacency and Prefix SIDs, then PCE MAY switch between various representations of same path.

If P flag is cleared, the PCE MAY recompute if current path is not considered valid, for example after topology update resulting in path not satisfying LSP’s path constraints, but it MUST NOT recompute path if current path is not optimal.

If P flag is set, the PCE MUST NOT recompute path during LSP lifetime even if path is invalidated. Only exception is explicit request from operator to recompute path

If F flag is cleared, path update triggered manually by operator or any northbound interface of PCE MAY be done. If flag is set the PCE CAN update path only to tear down LSP by sending PCUpdate message with empty ERO.

TLV MAY be included in PCInitiate and PCUpdate messages to indicate, which triggers will be disabled on the PCE. PCC should reflect flag values in PCRpt messages to forward requirement to other PCEs in the network.

5. Manageability Considerations

All manageability requirements and considerations listed in [RFC5440], [RFC8231] and [RFC8281] apply to PCEP protocol extensions defined in this document. In addition, requirements and considerations listed in this section apply.

5.1. Control of Function and Policy

A PCE or PCC implementation MAY allow the capability of supporting PCEP extensions introduced in this document to be enabled/disabled as part of the global configuration.

5.2. Information and Data Models

An implementation SHOULD allow the operator to view the capability defined in this document. Section 4.1 and 4.1.1 of [I-D.ietf-pce-pcep-yang] should be extended to include that capability for PCEP peer.

Section 4.2 of [I-D.ietf-pce-pcep-yang] module should be extended to add notification for blocked recomputation satisfying specified constraints if recomputation is blocked using PATH-RECOMPUTATION TLV.

5.3. Liveness Detection and Monitoring

Circuit-Style Policy draft [I-D.ietf-spring-cs-sr-policy] is already describing connectivity verification and path validity considerations for Circuit Style Policies.

5.4. Verify Correct Operations

A PCE implementation SHOULD notify operator in case blocked recomputation for path, which is no longer satisfying specified constraints and it SHOULD allow the operator to view LSPs on PCE, which are not satisfying specified constraints.

5.5. Requirements On Other Protocols

The PCEP extensions defined in this document do not imply any new requirements on other protocols. Overall concept of Circuit Style policies requires interaction with other protocols, but those requirements are already described in [I-D.ietf-spring-cs-sr-policy].

5.6. Impact On Network Operations

The mechanisms defined in [RFC5440], [RFC8231], and [RFC8281] also apply to the PCEP extensions defined in this document.

6. Implementation Status

[Note to the RFC Editor - remove this section before publication, as well as remove the reference to RFC 7942.]

This section records the status of known implementations of the protocol defined by this specification at the time of posting of this Internet-Draft, and is based on a proposal described in [RFC7942]. The description of implementations in this section is intended to assist the IETF in its decision processes in progressing drafts to RFCs. Please note that the listing of any individual implementation here does not imply endorsement by the IETF. Furthermore, no effort has been spent to verify the information presented here that was supplied by IETF contributors. This is not intended as, and must not be construed to be, a catalog of available implementations or their features. Readers are advised to note that other implementations may exist.

According to [RFC7942], "this will allow reviewers and working groups to assign due consideration to documents that have the benefit of running code, which may serve as evidence of valuable experimentation and feedback that have made the implemented protocols more mature. It is up to the individual working groups to use this information as they see fit".

6.1. Cisco

* Organization: Cisco Systems

* Implementation: IOS-XR PCC and PCE.

* Description: PCEP extensions supported using VENDOR-INFORMATION Object.

* Maturity Level: Production.

* Coverage: Partial.

* Contact: ssidor@cisco.com

7. Security Considerations

The security considerations described in [RFC5440], [RFC8231], [RFC8253],[RFC8281] and [RFC8664] in itself.

Note that this specification introduces possibility to block path recomputation after various topology events. This creates an additional vulnerability if the security mechanisms of [RFC5440], [RFC8231], and [RFC8281] are not used. If there is no integrity protection on the session, then an attacker could block path updates from PCE potentially resulting in traffic drop.

8. IANA Considerations

8.1. LSP-EXTENDED-FLAG TLV

[RFC9357] defines the LSP-EXTENDED-FLAG TLV. IANA is requested to make the following assignment from the "LSP-EXTENDED-FLAG TLV Flag Field" registry:

Bit	Description	Reference
TBD1	Strict-Path Flag (O)	This document

Table 1

8.2. PATH-RECOMPUTATION TLV

IANA is requested to make the assignment of a new value for the existing "PCEP TLV Type Indicators" registry as follows:

TLV Type	TLV Name	Reference
TBD2	PATH-RECOMPUTATION TLV	This document

Table 2

9. References

9.1. Normative References

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <https://www.rfc-editor.org/info/rfc2119>.

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

[RFC8174] Leiba, B., "Ambiguity of Uppercase vs Lowercase in RFC 2119 Key Words", BCP 14, RFC 8174, DOI 10.17487/RFC8174, May 2017, <https://www.rfc-editor.org/info/rfc8174>.

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

[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>.

[RFC8664] Sivabalan, S., Filsfils, C., Tantsura, J., Henderickx, W., and J. Hardwick, "Path Computation Element Communication Protocol (PCEP) Extensions for Segment Routing", RFC 8664, DOI 10.17487/RFC8664, December 2019, <https://www.rfc-editor.org/info/rfc8664>.

[RFC9357] Xiong, Q., "Label Switched Path (LSP) Object Flag Extension for Stateful PCE", RFC 9357, February 2023, <https://www.rfc-editor.org/info/rfc9357>.

9.2. 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-22, 31 October 2019, <https://datatracker.ietf.org/doc/html/draft-ietf-pce-pcep-yang-22>.

[I-D.ietf-spring-cs-sr-policy]
Schmutzer, C., Ali, Z., Maheshwari, P., Rokui, R., and A. Stone, "Circuit Style Segment Routing Policies", Work in Progress, Internet-Draft, draft-ietf-spring-cs-sr-policy-01, 16 June 2023, <https://www.ietf.org/archive/id/draft-ietf-spring-cs-sr-policy-01.txt>.

[RFC7942] Sheffer, Y. and A. Farrel, "Improving Awareness of Running Code: The Implementation Status Section", BCP 205, RFC 7942, DOI 10.17487/RFC7942, July 2016, <https://www.rfc-editor.org/info/rfc7942>.

[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>.

[SDH] International Telecommunication Union, "Network node interface for the synchronous digital hierarchy (SDH)", ITU-T Recommendation G.707, October 2020.

[SONET] American National Standards Institute, "Synchronous Optical Network (SONET) Basic Description including Multiplex Structure, Rates, and Formats", ANSI T1.105, January 1995.

Contributors

Daniel Voyer
Bell Canada
Email: daniel.voyer@bell.ca

Reza Rokui
Ciena
Email: rrokui@ciena.com

Tarek Saad
Cisco Systems, Inc.
Email: tsaad.net@gmail.com

Zafar Ali
Cisco Systems, Inc.
Email: zali@cisco.com

Authors' Addresses

Samuel Sidor
Cisco Systems, Inc.
Eurovea Central 3.
811 09 Bratislava
Slovakia
Email: ssidor@cisco.com

Praveen Maheshwari
Airtel India
Email: Praveen.Maheshwari@airtel.com

Andrew Stone
Nokia
Email: andrew.stone@nokia.com

Luay Jalil
Verizon
Email: luay.jalil@verizon.com

Shuping Peng
Huawei Technologies
Email: pengshuping@huawei.com

PCE Working Group
Internet-Draft
Intended status: Standards Track
Expires: September 4, 2022

Y. Wang
A. Wang
China Telecom
F. Qin
China Mobile
H. Chen
Futurewei
C. Zhu
ZTE Corporation
March 3, 2022

PCEP Procedures and Extension for VLAN-based Traffic Forwarding
draft-wang-pce-vlan-based-traffic-forwarding-05

Abstract

This document defines the Path Computation Element Communication Protocol (PCEP) extension for VLAN-based traffic forwarding in native IP network and describes the essential elements and key processes of the data packet forwarding system based on VLAN info to accomplish the End to End (E2E) traffic assurance for VLAN-based traffic forwarding in native IP network.

Status of This Memo

Copyright Notice

Table of Contents

1. Introduction

[RFC8283] introduces the architecture for the PCE as a central controller as an extension to the architecture described in [RFC4655]. Based on such mechanism, the PCE can calculate the optimal path for various applications and send the instructions to the network equipment via PCEP protocol, thus control the packet

forwarding and achive the QoS assurance effects for priority traffic.
.

[RFC8735] describes the scenarios of QoS assurance for hybrid cloud-based application within one domain and traffic engineering in multi-domain. It proposes also the consideration for the potential solution, that is:

1. Should be applied both in native IPv4 and IPv6 environment.

2. Should be same procedures for the intra-domain and inter-domain scenario.

3. Should utilize the existing forwarding capabilities of the deployed network devices.

With the large scale deployment of Ethernet interfaces in operator network and PCECC architecture, it is possible to utilize the VLAN information within the Ethernet header to build one end-to-end dedicated path to guide the forwarding of the packet. Similar with the PCECC for LSP [RFC9050], this document defines a Path Computation Element Communication Protocol (PCEP) Extension for VLAN-based traffic forwarding by using the VLAN info contained in the Ethernet frame in native IP network and the mechanism is actually the PCECC for VSP(VLAN Switching Path). It is an end to end traffic guarantee mechanism based on the PCEP protocol in the native IP environment, which can ensure the connection-oriented network communication. It can simplify the calculation and forwarding process of the optimal path by blending it with elements of PCEP and without necessarily completely replacing it. The overall QoS assurance effect is achieved via the central controller by calculating and deploying the optimal VSP to bypass the congested nodes and links, thus avoids the resource reservation on each nodes in advance.

Compared with other traffic assurance technologies such as MPLS or srv6 which is supported only in IPv6 environment, and has the obvious packet overhead problems, the VLAN-based traffic forwarding (VTF) mechanism uses a completely new address space which will not conflict with other existing protocols and can easily avoid these problems and be deployed in IPv4 and IPv6 environment simultaneously. It is suitable for ipv4 and ipv6 networks and can leverage the existing PCE technologies as much as possible.

2. Conventions used in this document

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

3. Terminology

The following terms are defined in this draft:

- o PCC: Path Computation Client
- o PCE: Path Computation Element
- o PCEP: PCE Communication Protocol
- o PCECC: PCE-based Central Controller
- o LSP: Lable Switching Path
- o PST: Path Setup Type

4. Procedures for VLAN-based Traffic Forwarding

The target deployment environment of VLAN based traffic forwarding mechanism is for Native IP(IPv4 and IPv6). In such scenarios, the BGP is used for the prefix distribution among underlying devices(PCCs), no MPLS is involved.

In order to set up the VLAN-based traffic forwarding paths for different applications in native IP network, multiple BGP sessions should be deployed between the ingress PCC and egress PCC at the edge of the network respectively.

Based on the business requirements, the PCE calculates the explicit route and sends the route information to the PCCs through PCInitiate messages. When received the PCInitiate message, the ingress PCC will form a VLAN-Forwarding routing table defined in this document. The packet to be guaranteed will be matched in the table and then be labeled with corresponding VLAN tag. The labeled packet will be further sent to the PCC’s specific subinterface identified by the VLAN tag and then be forwarded. Similarly, the transit PCC and the egress PCC will form a VLAN-Crossing routing table after received the PCInitiate message. The packet to be guaranteed will be relabled with new VLAN tag and then be forwarded. For PCC, there is no corresponding VLAN allocation mechanism at present which is different with the label in MPLS, so the mechanism of allocating and managing VLAN ID by PCC will not be considered in this draft as per [RFC9050].

The whole procedures mainly focus on the end-to-end traffic for key application which can ensure the adequacy of VLAN number for this scenario. During the whole packet forwarding process, the packet can be encapsulated with reserved multicast MAC addresses(e.g.

0180:C200:0014 for ISIS levle1, 0180:C200:0015 for ISIS levle2) and don’t need to change hop by hop so as to accept by each PCC.

5. Capability Advertisement

During the PCEP Initialization Phase, PCEP Speakers (PCE or PCC) advertise their support of VLAN-based trafficforwarding extensions. This document defines a new Path Setup Type (PST)[RFC8408] for PCECC, as follows:

- o PST=TBD1: Path is a VLAN-based traffic forwarding type.

A PCEP speaker MUST indicate its support of the function described in this document by sending a PATH-SETUP-TYPE-CAPABILITY TLV in the OPEN object with this new PST included in the PST list.

Because the path is set up through PCE, a PCEP speaker must advertise the PCECC capability by using PCECC-CAPABILITY sub-TLV which is used to exchange information about their PCECC capability as per PCEP extensions defined in [RFC9050]

A new flag is defined in PCECC-CAPABILITY sub-TLV for VLAN-based traffic forwarding.

V (VLAN-based-forwarding-CAPABILITY - 1 bit - TBD2): If set to 1 by a PCEP speaker, it indicates that the PCEP speaker supports the capability of VLAN based traffic forwarding as specified in this document. The flag MUST be set by both the PCC and PCE in order to support this extension.

If a PCEP speaker receives the PATH-SETUP-TYPE-CAPABILITY TLV with the newly defined path setup type, but without the V bit set in PCECC-CAPABILITY sub-TLV, it MUST:

- o Send a PCErr message with Error-Type=10(Reception of an invalid object) and Error-Value TBD3(PCECC VLAN-based-forwarding-CAPABILITY bit is not set).
- o Terminate the PCEP session

6. PCEP message

As per [RFC8281] ,the PCInitiate message sent by a PCE was defined to trigger LSP instantiation or deletion with the SRP and LSP object included during the PCEP initialization phase. The Path Computation LSP State Report message (PCRpt message) was defined in [RFC8231], which is used to report the current state of a LSP. A PCC can send a LSP State Report message in response to a LSP instantiation.

Besides, the message can either in response to a LSP Update Request from a PCE or asynchronously when the state of a LSP changes .

[RFC9050] defines an object called Central Controller Instructions (CCI) to specify the forwarding instructions to the PCC. During the coding process used for central controller instructions, the object contains the label information and is carried within PCInitiate or PCRpt message for label download .

This document specify two new CCI object-types for VLAN-based traffic forwarding in the native IP network and are said to be mandatory in a PCEP message when the object must be included for the message to be considered valid. In addition, this document extends the PCEP message to handle the VLAN-based traffic forwarding path in the native IP network with the new CCI object.

6.1. The PCInitiate message

The PCInitiate message[RFC8281] extended in[RFC9050] can be used to download or remove labels by using the CCI Object.

Based on the extended PCInitiate message and PCRpt described in [I-D.ietf-pce-pcep-extension-native-ip], the (BGP Peer Info (BPI) Object and the Peer Prefix Association (PPA) Object is used to establish multi BGP sessions and advertise route prefixes among different BGP sessions before setting up a VLAN-based traffic forwarding path.

This document extends the PCInitiate message as shown below:

```
<PCInitiate Message> ::= <Common Header>
                              <PCE-initiated-lsp-list>
   Where:
      <Common Header> is defined in [RFC5440]

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

      <PCE-initiated-lsp-request> ::=
                            (<PCE-initiated-lsp-instantiation>|
                             <PCE-initiated-lsp-deletion>|
                             <PCE-initiated-lsp-central-control>)

      <PCE-initiated-lsp-central-control> ::= <SRP>
                                              <LSP>
                                              <cci-list>|
                                              ((<BPI>|<PPA>)
                                              <new-CCI>)

      <cci-list> ::=  <new-CCI>
                      [<cci-list>]
```

Where:

<cci-list> is as per [RFC9050].
<PCE-initiated-lsp-instantiation> and <PCE-initiated-lsp-deletion> are as per [RFC8281].
<BPI> and <PPA> are as per [draft-ietf-pce-pcep-extension-native-ip-09]

When PCInitiate message is used to create VLAN-based forwarding instructions, the SRP, LSP and CCI objects MUST be present. The error handling for missing SRP, LSP or CCI object is as per [RFC9050]. Further only one of BPI, PPA or one type of CCI objects MUST be present. If none of them are present, the receiving PCE MUST send a PCErr message with Error- type=6 (Mandatory Object missing) and Error-value=TBD4 (VLAN-based forwarding object missing). If there are more than one of BPI, PPA or one type of CCI objects are presented, the receiving PCC MUST send a PCErr message with Error-type=19(Invalid Operation) and Error- value=TBD5(Only one of BPI, PPA or one type of the CCI objects for VLAN can be included in this message).

6.2. The PCRpt message

The PCRpt message is used to report the state and confirm the VLAN info that were allocated by the PCE, to be used during the state synchronization phase or as acknowledgement to PCInitiate message.

The format of the PCRpt message is as follows:

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

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

         <state-report> ::= (<lsp-state-report>|
                            <central-control-report>)

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

         <central-control-report> ::= [<SRP>]
                                      <LSP>
                                      <cci-list>|
                                      ((<BPI>|<PPA>)
                                      (<new-CCI>)


        Where:
         <path> is as per [RFC8231] and the LSP and SRP object are
         also defined in [RFC8231].
         <BPI> and <PPA> are as per
         [draft-ietf-pce-pcep-extension-native-ip-09]
```

The error handling for missing LSP or CCI object is as per [RFC9050]. Further only one of BPI, PPA or one type of CCI objects MUST be present. If none of them are present, the receiving PCE MUST send a PCErr message with Error- type=6 (Mandatory Object missing) and Error-value=TBD4 (VLAN-based forwarding object missing). If there are more than one of BPI, PPA or one type of CCI objects are presented, the receiving PCC MUST send a PCErr message with Error-type=19(Invalid Operation) and Error- value=TBD5(Only one of BPI, PPA or one type of the CCI objects for VLAN can be included in this message).

7. VSP Operations

Based on [RFC8281] and [RFC9050], in order to set up a PCE-initiated VSP based on the PCECC mechanism, a PCE needs to send a PCInitiate message with the PST set to TBD1 in SRP for the PCECC to the ingress PCC.

The VLAN-forwarding instructions from the PCECC needs to be sent after the initial PCInitiate and PCRpt message exchange with the

ingress PCC. On receipt of a PCInitiate message for the PCECC VSP, the PCC responds with a PCRpt message with the status set to 'Going-up', carrying the assigned PLSP-ID and set the D(Delegate) flag and C(Create) flag(see Figure 1).

After that, the PCE needs to send a PCInitiate message to each node along the path to download the VLAN instructions. The new CCI for the VLAN operations in PCEP are done via the PCInitiate message by defining a new PCEP object for CCI operations. The LSP and the LSP-IDENTIFIERS TLV are described for the RSVP-signaled LSPs but are applicable to the PCECC VSP as well. So the LSP is included in the PCInitiate message can still be used to identify the PCECC VSP for this instruction and the process is the same.

When the PCE receives this PCRpt message with the PLSP-ID, it assigns VLAN along the path and sets up the path by sending a PCInitiate message to each node along the path of the VSP, as per the PCECC technique. The ingress PCC would receive one VLAN forwarding CCI Object which contains VLAN on the logical subinterface and the Peer IP address. The transit PCC would receive two VLAN crossing CCI Objects with the O bit set for the out-VLAN on the egress subinterface and the O bit unset for the in-VLAN on the ingress subinterface. Similar with the transit PCC, the egress PCC would receive two VLAN crossing CCI Objects but the out-VLAN on the egress subinterface is set to 0. Once the VLAN operations are completed, the PCE MUST send a PCUpd message to the ingress PCC.

```
                  +-------+                              +-------+
                  |PCC    |                              |  PCE  |
                  |ingress|                              +-------+
          +------|       |                                  |
          | PCC  +-------+                                  |
          | transit| |                                     |
    +------|        | |<--PCInitiate,PLSP-ID=0,PST=TBD1------| PCECC VSP
    |PCC   +--------+ |----PCRpt,PLSP-ID=2,D=1,C=1---------->| Initiate
    |egress  |  |     |         (GOING-UP)                  | PCECC VSP
    +--------+  |     |                                     |
        |       |     |                                     |
        |<-------PCInitiate,VLAN-CROSSING-CC-ID=X1,X2--------| VLAN
        |       | PLSP-ID=2,IN-VLAN=N1,OUT-VLAN=0             | download
        |       |     |                                     | CCI
        |--------PCRpt,VLAN-CROSSING-CC-ID=X1,X2------------>|
        |       |PLSP-ID=2,IN-VLAN=N1,OUT-VLAN=0              |
        |       |     |                                     |
        |       |<---PCInitiate,VLAN-CROSSING-CC-ID=Y1,Y2--->| VLAN
        |       |     |PLSP-ID=2,IN-VLAN=N2,OUT-VLAN=N1       | download
        |       |     |                                     | CCI
        |       |-------PCRpt,VLAN-CROSSING-CC-ID=Y1,Y2----->|
        |       |     |PLSP-ID=2,IN-VLAN=N2,OUT-VLAN=N1       |
        |       |     |                                     |
        |       |     |<--PCInitiate,VLAN-FORWARDING-CC-ID=Z-| VLAN
        |       |     |          PLSP-ID=2,VLAN=N2            | download
        |       |     |                                     | CCI
        |       |     |-----PCRpt,CC-ID=Z,PLSP-ID=2--------->|
        |       |     |          PLSP-ID=2,VLAN=N2            |
        |       |     |                                     |
        |       |     |<---PCUpd,PLSP-ID=2,PST=TBD1,D=1------| PCECC VSP
        |       |     |      (UP)                           | Update
        |       |     |----PCRpt,PLSP-ID=2,D=1,C=1---------->|
        |       |     |      (UP)                           |
```

Figure 1: PCE-Initiated PCECC VSP

In order to delete an LSP based on the PCECC, the PCE sends CCI and SRP object with the R bit set to 1 via a PCInitiate message to each node along the path of the VSP to clean up the label-forwarding instruction.

As per [RFC9050], the PCECC VSP also follows the same make-before-break principles. As shown in the figure 2, new path for VSP triggers the new CCI Distribution process. The PCECC first updates the new VLAN instructions and informs each node along the new path through the new VLAN crossing CCI Objects and VLAN forwarding CCI Objects to download the new VSP. The PCUpd message then triggers the traffic switch on the updated path. On receipt of the PCRpt message corresponding to the PCUpd message, the PCE does the cleanup

operation for the former VSP,which is the same as the LSP update process.

```
                +-------+                              +-------+
                |PCC    |                              |  PCE  |
                |ingress|                              +-------+
         +------|       |                                  |
         | PCC  +-------+                                  |
         | transit| |                                     |
  +------|        | |                                     |
  |PCC   +--------+ |                                     |
  |egress  |  |     |                                     |
  +--------+  |     |                                     |
      |       |     |                                     |
      |<----- PCInitiate,VLAN-CROSSING-CC-ID=NEW-X1,X2----| New Path
      |       |   PLSP-ID=1,IN-VLAN=NEW-N1,OUT-VLAN=0     | for VSP
      |       |     |                                     | triggers
      |--------PCRpt,VLAN-CROSSING-CC-ID=NEW-X1,X2------->| new CCI
      |       | PLSP-ID=1,IN-VLAN=NEW-N1,OUT-VLAN=0       |
      |       |     |                                     |
      |       |<----------PCInitiate,PLSP-ID=1------------|
      |       |     |VLAN-CROSSING-CC-ID=NEW-Y1,NEW-Y2    | Label
      |       |     |  IN-VLAN=NEW-N2,OUT-VLAN=NEW-N1     | download
      |       |     |                                     | CCI
      |       |--------------PCRpt,PLSP-ID=1------------->|
      |       |     |VLAN-CROSSING-CC-ID=NEW-Y1,NEW-Y2    |
      |       |     |  IN-VLAN=NEW-N2,OUT-VLAN=NEW-N1     |
      |       |     |                                     |
      |       |     |<--------PCInitiate,PLSP-ID=1--------| Label
      |       |     |     VLAN-FORWARDING-CC-ID=NEW-Z     | download
      |       |     |            VLAN=NEW-N2              | CCI
      |       |     |                                     |
      |       |     |----------PCRpt,PLSP-ID=1----------->|
      |       |     |     LAN-FORWARDING-CC-ID=NEW-Z      |
      |       |     |           VLAN=NEW-N2               |
      |       |     |                                     |
      |       |     |<---PCUpd,PLSP-ID=1,PST=TBD1,D=1-----| PCECC
      |       |     |    (SRP=S)                          | VSP Update
      |       |     |                                     |
      |       |     |---PCRpt,PLSP-ID=1,PST=TBD1,D=1----->| Trigger
      |       |     |       (SRP=S)                       | Delete
      |       |     |                                     | former CCI
      |       |     |                                     |
      |<--------------PCInitiate, PLSP-ID=1---------------| Label
      |       |     |VLAN-CROSSING-CC-ID=X1,X2,R=1        | cleanup
      |----------------PCRpt,PLSP-ID=1------------------->| CCI
      |       |     |VLAN-CROSSING-CC-ID=X1,X2,R=1        |
      |       |     |                                     |
```

```
        |        |<------------PCInitiate,PLSP-ID=1----------| Label
        |        |     |VLAN-CROSSING-CC-ID=Y1,Y2,R=1         | cleanup
        |        |---------------PCRpt,PLSP-ID=1------------>| CCI
        |        |     |VLAN-CROSSING-CC-ID=Y1,Y2,R=1         |
        |        |     |                                      |
        |        |     |<--------PCInitiate,PLSP-ID=1--------| Label
        |        |     |VLAN-FORWARDING-CC-ID=Z,R=1           | cleanup
        |        |     |---------PCRpt,PLSP-ID=1------------>| CCI
        |        |     |VLAN-FORWARDING-CC-ID=Z,R=1           |
```

Figure 2: PCECC VSP Update

8. VXLAN-based traffic forwarding Procedures

8.1. Multiple BGP Session Establishment Procedures

As described in section 4, multiple BGP sessions should be deployed between the ingress device and egress device at the edge of the network respectively in order to carry information of different applications. As per [I-D.ietf-pce-pcep-extension-native-ip], the PCE should send the BPI((BGP Peer Info) Object to the ingress and egress device with the indicated Peer AS and Local/Peer IP address. The Ingress and egress devices will receive multiple BPI objects to establish sessions with different next hop. The specific process is as follows:

```
             +----------------------+
   +---------+-        PCE          + --------+
   |         +----------^-----------+         |
   |          |         |          |          |
   |        +--+       +--+       +--+        |
   |------- +R2+ ------+R3+-------+R4+ --------
   |        +--+       +--+       +--+        |
   |                                          |
   +--+                +--+                +--+
   +R1+----------------+R5+----------------+R6+
   +--+                +--+                +--+
   |                                          |
   |<------------- BGP Session A ------------>|
   |<------------- BGP Session B ------------>|
   |<------------- BGP Session C ------------>|
```

Figure 3: BGP Session Establishment Procedures

8.2. BGP Prefix Advertisement Procedures

The detail procedures for BGP prefix advertisement procedures is introduced in [I-D.ietf-pce-pcep-extension-native-ip], using PCInitiate and PCRpt message pair.

The BGP prefix for different BGP sessions should be sent to the ingress and egress device respectively. The end-to-end traffic for key application can be identified based on these BGP prefix informations and be further assured. As per [I-D.ietf-pce-pcep-extension-native-ip], the PPA(Peer Prefix Association) object with list of prefix subobjects and the peer address will be sent through the PCInitiate and PCRpt message pair. Through BGP protocol, the ingress device can learn different BGP prefix of the egress device based on the different BGP sessions.

8.3. VLAN mapping info Advertisement Procedures

After the BGP prefix for different BGP session are successfully advertised, information of different applications should be forwarded to different VLAN-based traffic forwarding paths. In order to set up a VLAN-based traffic forwarding path, the PCE should send the VLAN forwarding CCI Object with the VLAN-ID included to the ingress PCC and the VLAN crossing CCI Object to the transit PCC and egress PCC.

8.3.1. VLAN-Based forwarding info Advertisement Procedures

The detail procedures for VLAN-Based forwarding info advertisement contained in the VLAN forwarding CCI Object is shown below, using PCInitiate and PCRpt message pair.

The VLAN forwarding CCI Object should be sent through the PCInitiate and PCRpt message pair. After the PCC receives the CCI object (with the R bit set to 0 in SRP object) in PCInitiate message, the PCC will form a VLAN-Forwarding routing table and the PCC’s subinterface will set up the specific VLAN based on the VLAN forwarding CCI object, source and destination BGP prefix learnt before. When the ingress PCC receives a packet, it will look up the VLAN-Forwarding routing table based on the source and destination IP contained in the packet. The packet to be guaranteed will be matched in the table and then be labeled with corresponding VLAN tag. After that, The labeled packet will be further forwarded to the specific subinterface.

When PCC receives the VLAN forwarding CCI Object with the R bit set to 1 in SRP object in PCInitiate message, the PCC should withdraw the VLAN-Based forwarding info advertisement to the peer that indicated by this object.

On receipt of a PCInitiate message for the PCECC VSP, the PCC should report the result via the PCRpt messages, with the corresponding SRP and CCI object included.

```
             +----------------------+
   +---------+         PCE          + --------+
   |         +----------^-----------+         |
   |          |          |          |          |
  M1&M1-R     |          |          |          |
   |          |          |          |          |
   |          |          |          |          |
   |         +--+       +--+       +--+        |
   |------- +R2+ ------+R3+-------+R4+ --------
   |         +--+       +--+       +--+        |
   |                                           |
   +--+                +--+                +--+
   +R1+----------------+R5+----------------+R6+
   +--+                +--+                +--+
```

Figure 4: VLAN-Based forwarding info Advertisement Procedures for Ingress PCC

The message number, message peers, message type and message key parameters in the above figures are shown in below table:

Table 1: Message Information

No.	Peers	Type	Message Key Parameters
M1 M1-R	PCE/R1	PCInitiate PCRpt	CC-ID=X1(Symbolic Path Name=Class A) VLAN Forwarding CCI Object (Peer_IP=R6_A,Interface_Address=INF1, VLAN_ID=VLAN_R1_R2)

VLAN-Forwarding routing table maintained in the ingress PCC is as follows, which is used to match the packet to be guaranteed based on the source and destination BGP prefix.

Table 2: VLAN-Forwarding routing table

Dst IP Address	Interface	VLAN
Prefixes from R6 Session1	INF 1	VLAN_R1_R2
Prefixes from R6 SessionX	INF X	X
...		

8.3.2. VLAN-Based crossing info Advertisement Procedures

The detail procedures for VLAN-Based crossing info advertisement contained in the VLAN crossing CCI Object is shown below, using PCInitiate and PCRpt message pair.

The PCC would receive VLAN crossing CCI Objects with the in-VLAN CCI without the O bit set and the out-VLAN CCI with the O bit set. After the process of VLAN-Based forwarding info advertisement mentioned above, the PCC will form a VLAN-crossing routing table and the PCC's subinterface will set up the specific VLAN based on the VLAN crossing CCI Object(with the R bit set to 0 in SRP object) contained in the PCInitiate message. The VLAN-crossing routing table consists of an in-VLAN tag and an out-VLAN tag which specifies a new VLAN forwarding path. When the transit PCC receives a data packet that has been labeled with VLAN by ingress PCC before, it will look up the VLAN-Crossing routing table based on the VLAN tag. If matched, the in-VLAN tag of this data packet will be replaced by a new out-VLAN tag of the current transit PCC according to the table. The packet with the new VLAN tag will be further forwarded to the next hop.

For the egress PCC, the out-VLAN tag in the VLAN-crossing routing table should be 0 which indicates it is the last hop of the transmission. So the egress PCC will directly remove the in-VLAN tag of the packet and the packet will be forwarded.

When PCC receives the VLAN crossing CCI Object with the R bit set to 1 in SRP object in PCInitiate message, the PCC should withdraw the VLAN-Based crossing info advertisement to the peer that indicated by this object.

On receipt of a PCInitiate message for the PCECC VSP, the PCC should report the result via the PCRpt messages, with the corresponding SRP and CCI object included.

When the out-VLAN tag conflicts with a pre-defined VLAN tag or the PCC can not set up a VLAN forwarding path with the out-VLAN tag, an error (Error-type=TBD6, VLAN-based forwarding failure, Error-value=TBD7, VLAN crossing CCI Object peer info mismatch) should be reported via the PCRpt message.

```
                  +----------------------+
         +---------+         PCE          + --------+
         |         +----------^-----------+         |
         |          |         |          |          |
         |       M1&M1-R    M2&M2-R   M3&M3-R    M4&M4-R
         |          |         |          |          |
         |        +--+       +--+       +--+        |
         |------- +R2+ ------+R3+-------+R4+ -------|
         |        +--+       +--+       +--+        |
         |                                          |
         +--+                +--+                +--+
         +R1+----------------+R5+----------------+R6+
         +--+                +--+                +--+
```

Figure 5: VLAN-Based crossing info Advertisement Procedures for transit PCC and egress PCC

The message number, message peers, message type and message key parameters in the above figures are shown in below table:

Table 3: Message Information

No.	Peers	Type	Message Key Parameters
M1 M1-R	PCE/R2	PCInitiate PCRpt	CC-ID=X1(Symbolic Path Name=Class A) VLAN crossing CCI Object(IN) (O=0,Interface_Address=INF1,IN_VLAN_ID=VLAN_R1_R2) VLAN crossing CCI Object(OUT) (O=1,Interface_Address=INF2,OUT_VLAN_ID=VLAN_R2_R3)
M2 M2-R	PCE/R3	PCInitiate PCRpt	CC-ID=X1(Symbolic Path Name=Class A) VLAN crossing CCI Object(IN) (O=0,Interface_Address=INF1,IN_VLAN_ID=VLAN_R2_R3) VLAN crossing CCI Object(OUT) (O=1,Interface_Address=INF2,OUT_VLAN_ID=VLAN_R3_R4)
M3 M3-R	PCE/R4	PCInitiate PCRpt	CC-ID=X1(Symbolic Path Name=Class A) VLAN crossing CCI Object(IN) (O=0,Interface_Address=INF1,IN_VLAN_ID=VLAN_R3_R4) VLAN crossing CCI Object(OUT) (O=1,Interface_Address=INF2,OUT_VLAN_ID=VLAN_R4_R6)
M4 M4-R	PCE/R6	PCInitiate PCRpt	CC-ID=X1(Symbolic Path Name=Class A) VLAN crossing CCI Object(IN) (O=0,Interface_Address=INF1,IN_VLAN_ID=VLAN_R4_R6) VLAN crossing CCI Object(OUT) (O=1,Interface_Address=INF2,OUT_VLAN_ID=0)

VLAN-Crossing routing table maintained in the transit PCC and egress PCC is as follows. Through the mapping of the in-VLAN and the out VLAN, the data packet to be guaranteed will be transferred to the specific interface and be switched on the out VLAN for the transit PCC or 0 for the egress PCC.

Table 4: VLAN-Crossing routing table

IN-Interface	IN-VLAN	OUT-Interface	OUT-VLAN
INF1	VLAN_R1_R2	INF2	VLAN_R2_R3
INF3	X	INF4	Y
INF5	Z	INF6	0
	...		

9. New PCEP Objects

The Central Control Instructions (CCI) Object is used by the PCE to specify the forwarding instructions is defined in [RFC9050]. This document defines another two CCI object-types for VLAN-based traffic forwarding network. All new PCEP objects are compliant with the PCEP object format defined in [RFC5440].

9.1. VLAN forwarding CCI Object

The VLAN forwarding CCI Object is used to set up the specific VLAN forwarding path of the logical subinterface that the traffic will be forwarded to and transfer the packet to the specific hop. Combined with this type of CCI Object and the Peer Prefix Association object(PPA) defined in [I-D.ietf-pce-pcep-extension-native-ip], the ingress PCC will form a VLAN-Forwarding routing table which is used to identify the traffic that needs to be protected. This object should only be included and sent to the ingress PCC of the end2end path.

CCI Object-Class is 44.

CCI Object-Type is TBD8 for VLAN forwarding info in the native IP network.

```
 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
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                            CC-ID                              |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|            Reserved1           |            Flags              |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|       VLAN-ID          |             Reserved2                |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                                                               |
//                     Interface Address TLV                    //
|                                                               |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                                                               |
//                     Peer IP Address TLV                      //
|                                                               |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                                                               |
//                         Additional TLVs                      //
|                                                               |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 6: VLAN Forwarding CCI Object

The fields in the CCI object are as follows:

CC-ID: is as described in [RFC9050]. Following fields are defined for CCI Object-Type TBD8.

Reserved1(16 bits): is set to zero while sending, ignored on receipt.

Flags(16 bits): is used to carry any additional information pertaining to the CCI. Currently no flag bits are defined.

VLAN ID(12 bits):the ID of the VLAN forwarding path that the PCC will set up on its logical subinterface in order to transfer the packet to the specific hop.

Reserved2(20 bits): is set to zero while sending, ignored on receipt.

Interface Address TLV [RFC8779] MUST be included in this CCI Object-Type TBD8 to specify the interface which will set up the vlan defined in the VLAN Forwarding CCI Object.

The Peer IP Address TLV[RFC8779]MUST be included in this CCI Object-Type TBD8 to identify the end to end TE path in VLAN-based traffic forwarding network and MUST be unique.

9.2. Address TLVs

[RFC8779] defines IPV4-ADDRESS, IPV6-ADDRESS, and UNNUMBERED-ENDPOINT TLVs for the use of Generalized Endpoint. The same TLVs can also be used in the CCI object to find the Peer address that matches egress PCC and further identify the packet to be guaranteed. If the PCC is not able to resolve the peer information or can not find the corresponding ingress device, it MUST reject the CCI and respond with a PCErr message with Error-Type = TBD6 ("VLAN-based forwarding failure") and Error Value = TBD9 ("Invalid egress PCC information").

9.3. VLAN crossing CCI Object

The VLAN crossing CCI object is defined to control the transmission-path of the packet by VLAN-ID. This new type of CCI Object can be carried within a PCInitiate message sent by the PCE to the transit PCC and the egress PCC in the VLAN-based traffic forwarding scenarios.

CCI Object-Class is 44.

CCI Object-Type is TBD10 for VLAN crossing info in the native IP network.

```
 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
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                            CC-ID                              |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|            Reserved1           |            Flags           |O|
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|     VLAN-ID(in/out)    |             Reserved2               |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                                                               |
//                     Interface Address TLV                    //
|                                                               |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                                                               |
//                        Additional TLVs                       //
|                                                               |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 7: VLAN Crossing CCI Object

CC-ID: is as described in [RFC9050]. Following fields are defined for CCI Object-Type TBD10.

Reserved1(16 bits): is set to zero while sending, ignored on receipt.

Flags(16 bits): is used to carry any additional information pertaining to the CCI. Currently, the following flag bit are defined:

* O bit (out-label) : If the bit is set to '1', it specifies the VLAN is the out-VLAN, and it is mandatory to encode the egress interface information(via Interface Address TLVs in the CCI object). If the bit is not set or set to '0', it specifies the VLAN is the in-VLAN, and it is mandatory to encode the ingress interface information.

VLAN ID(12 bits): The ID of the VLAN switching path. When the O bit is set to 0, the VLAN is the in-VLAN and the ID indicates a VLAN forwarding path which is used to identify the traffic that needs to be protected. When the O bit is set to 1, the VLAN is the out-VLAN and it indicates the ID of the VLAN forwarding path that the PCC will set up on its logical subinterface in order to transfer the packet labled with this VLAN ID to the specific hop. To the transit PCC, the value must not be 0 to indicate it is not the last hop of the VLAN-based traffic forwarding path. To the egress PCC, the value must be 0 to indicate it is the last hop of the VLAN-based traffic forwarding path.

Reserved2(8 bits): is set to zero while sending, ignored on receipt.

Interface Address TLV [RFC8779] MUST be included in this CCI Object-Type TBD8 to specify the interface which will set up the vlan defined in the VLAN Forwarding CCI Object.

10. Deployment Considerations

11. Security Considerations

12. IANA Considerations

12.1. Path Setup Type Registry

[RFC8408] created a sub-registry within the "Path Computation Element Protocol (PCEP) Numbers" registry called "PCEP Path Setup Types". IANA is requested to allocate a new code point within this registry, as follows:

Value	Description	Reference
TBD1	VLAN-Based Traffic Forwarding Path	This document

12.2. PCECC-CAPABILITY sub-TLV's Flag field

[RFC9050] created a sub- registry within the "Path Computation Element Protocol (PCEP) Numbers" registry to manage the value of the PCECC-CAPABILITY sub- TLV's 32-bits Flag field. IANA is requested to allocate a new bit position within this registry, as follows:

Value	Description	Reference
TBD2(V)	VLAN-Based Forwarding CAPABILITY	This document

12.3. PCEP Object Types

IANA is requested to allocate new registry for the PCEP Object Type:

Object-Class Value	Name	Reference
44	CCI Object-Type TBD8: VLAN forwarding CCI TBD10: VLAN crossing CCI	This document

12.4. PCEP-Error Object

IANA is requested to allocate new error types and error values within the "PCEP-ERROR Object Error Types and Values" sub-registry of the PCEP Numbers registry for the following errors:

Error-Type	Meaning	Error-value	Reference
6	Mandatory Object missing	TBD4:VLAN-based forwarding object missing	This document
10	Reception of an invalid object	TBD3:PCECC VLAN-based-forwarding -CAPABILITY bit is not set	This document
19	Invalid Operation	TBD5: Only one of BPI, PPA or one type of the CCI objects for VLAN can be included in this message	This document
TBD6	VLAN-based forwarding failure	TBD7: VLAN crossing CCI Object peer info mismatch	This document
		TBD9: Invalid egress PCC information	This document

13. Acknowledgement

14. Normative References

[I-D.ietf-pce-pcep-extension-for-pce-controller]
Li, Z., Peng, S., Negi, M. S., Zhao, Q., and C. Zhou, "Path Computation Element Communication Protocol (PCEP) Procedures and Extensions for Using the PCE as a Central Controller (PCECC) of LSPs", draft-ietf-pce-pcep-extension-for-pce-controller-14 (work in progress), March 2021.

[I-D.ietf-pce-pcep-extension-native-ip]
Wang, A., Khasanov, B., Fang, S., Tan, R., and C. Zhu, "PCEP Extension for Native IP Network", draft-ietf-pce-pcep-extension-native-ip-17 (work in progress), February 2022.

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <https://www.rfc-editor.org/info/rfc2119>.

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

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

[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>.

[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>.

[RFC8408] Sivabalan, S., Tantsura, J., Minei, I., Varga, R., and J. Hardwick, "Conveying Path Setup Type in PCE Communication Protocol (PCEP) Messages", RFC 8408, DOI 10.17487/RFC8408, July 2018, <https://www.rfc-editor.org/info/rfc8408>.

[RFC8735] Wang, A., Huang, X., Kou, C., Li, Z., and P. Mi, "Scenarios and Simulation Results of PCE in a Native IP Network", RFC 8735, DOI 10.17487/RFC8735, February 2020, <https://www.rfc-editor.org/info/rfc8735>.

[RFC8779] Margaria, C., Ed., Gonzalez de Dios, O., Ed., and F. Zhang, Ed., "Path Computation Element Communication Protocol (PCEP) Extensions for GMPLS", RFC 8779, DOI 10.17487/RFC8779, July 2020, <https://www.rfc-editor.org/info/rfc8779>.

[RFC9050] Li, Z., Peng, S., Negi, M., Zhao, Q., and C. Zhou, "Path Computation Element Communication Protocol (PCEP) Procedures and Extensions for Using the PCE as a Central Controller (PCECC) of LSPs", RFC 9050, DOI 10.17487/RFC9050, July 2021, <https://www.rfc-editor.org/info/rfc9050>.

Authors' Addresses

Yue Wang
China Telecom
Beiqijia Town, Changping District
Beijing, Beijing 102209
China

Email: wangy73@chinatelecom.cn

Aijun Wang
China Telecom
Beiqijia Town, Changping District
Beijing, Beijing 102209
China

Email: wangaj3@chinatelecom.cn

Fengwei Qin
China Mobile
32 Xuanwumenxi Ave.
Beijing 100032
China

Email: qinfengwei@chinamobile.com

Huaimo Chen
Futurewei
Boston
USA

Email: Huaimo.chen@futurewei.com

Chun Zhu
ZTE Corporation
50 Software Avenue, Yuhua District
Nanjing, Jiangsu 210012
China

Email: zhu.chun1@zte.com.cn

PCE Working Group
Internet-Draft
Intended status: Standards Track
Expires: 11 October 2024

Y. Wang
A. Wang
China Telecom
B. Khasanov
Yandex LLC
F. Qin
China Mobile
H. Chen
Futurewei
C. Zhu
ZTE Corporation
9 April 2024

PCEP Procedures and Extension for VLAN-based Traffic Forwarding
draft-wang-pce-vlan-based-traffic-forwarding-09

Abstract

This document defines the Path Computation Element Communication Protocol (PCEP) extension for VLAN-based traffic forwarding in native IP network and describes the essential elements and key processes of the data packet forwarding system based on VLAN info to accomplish the End to End (E2E) traffic assurance for VLAN-based traffic forwarding in native IP network.

Status of This Memo

Copyright Notice

Table of Contents

1. Introduction

[RFC8283] introduces the extension to the architecture described in [RFC4655] where PCE acts as a central controller and gets more responsibility for LSP provisioning on each hop. Based on such mechanism, the PCE can calculate the optimal path for various applications and send the instructions to the network equipment via PCEP protocol, thus control the packet forwarding and achieve the QoS assurance for prioritized traffic. .

[RFC8735] describes the scenarios of QoS assurance for hybrid cloud-based application within one domain and traffic engineering in multi-domains. It proposes also the following requirements for the potential solution:

1. Should be applied both in native IPv4 and IPv6 environment.

2. Should be same procedures for the intra-domain and inter-domain scenario.

3. Should utilize the existing forwarding capabilities of the deployed network devices.

Due to large scale of Ethernet interfaces used in operators network and need to establish P2P connectivity for them, an operator should currently use either VPWS or EVPN with MPLS signaling. This is not suitable for Native IP scenarios. Thus PCECC architecture can solve that problem for Native IP networks by building the end-to-end dedicated path based on a VLAN header to control the forwarding behavior of a packet. Similar with the PCECC for LSP [RFC9050], this document defines a Path Computation Element Communication Protocol (PCEP) Extension for VLAN-based traffic forwarding by using the VLAN info contained in the Ethernet frame in native IP network and the mechanism is actually the PCECC for VSP(VLAN Switching Path). It is an end to end traffic guarantee mechanism based on the PCEP protocol in the native IP environment, which can ensure the connection-oriented network communication. The overall QoS assurance effect is achieved via the central controller by calculating and deploying the optimal VSP to bypass the congested nodes and links, thus avoids the resource reservation on each nodes in advance.

Compared with other traffic assurance technologies such as MPLS or SRv6 which is supported only in IPv6 environment and has the obvious packet overhead problems, the VLAN-based traffic forwarding (VTF) mechanism uses a completely new address space which will not conflict with other existing protocols and can easily avoid these problems and be deployed in IPv4 and IPv6 environment simultaneously. It is suitable for IPv4 and IPv6 networks and can leverage the existing PCE technologies as much as possible.

2. Conventions used in this document

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

3. Terminology

The following terms are defined in this draft:

* PCC: Path Computation Client

* PCE: Path Computation Element

* PCEP: PCE Communication Protocol

* PCECC: PCE-based Central Controller

* LSP: Lable Switching Path

* PST: Path Setup Type

4. Capability Advertisement

During the PCEP Initialization Phase, PCEP Speakers (PCE or PCC) advertise their support of VLAN-based traffic forwarding extensions. This document defines a new Path Setup Type (PST)[RFC8408] for PCECC, as follows:

* PST=TBD1: Path is a VLAN-based traffic forwarding type.

A PCEP speaker MUST indicate its support of the function described in this document by sending a PATH-SETUP-TYPE-CAPABILITY TLV in the OPEN object with this new PST included in the PST list.

Because the path is set up through PCE, a PCEP speaker must advertise the PCECC capability by using PCECC-CAPABILITY sub-TLV which is used to exchange information about their PCECC capability as per PCEP extensions defined in [RFC9050]

A new flag is defined in PCECC-CAPABILITY sub-TLV for VLAN-based traffic forwarding.

V (VLAN-based-forwarding-CAPABILITY - 1 bit - TBD2): If set to 1 by a PCEP speaker, it indicates that the PCEP speaker supports the capability of VLAN based traffic forwarding as specified in this document. The flag MUST be set by both the PCC and PCE in order to support this extension.

If a PCEP speaker receives the PATH-SETUP-TYPE-CAPABILITY TLV with the newly defined path setup type, but without the V bit set in PCECC-CAPABILITY sub-TLV, it MUST:

* Send a PCErr message with Error-Type=10(Reception of an invalid object) and Error-Value TBD3(PCECC VLAN-based-forwarding-CAPABILITY bit is not set).

* Terminate the PCEP session

5. PCEP message

As per [RFC8281] ,the PCInitiate message sent by a PCE was defined to trigger LSP instantiation or deletion with the SRP and LSP object included during the PCEP initialization phase. The Path Computation LSP State Report message (PCRpt message) was defined in [RFC8231], which is used to report the current state of a LSP. A PCC can send a LSP State Report message in response to a LSP instantiation. Besides, the message can either in response to a LSP Update Request from a PCE or asynchronously when the state of a LSP changes .

[RFC9050] defines an object called Central Controller Instructions (CCI) to specify the forwarding instructions to the PCC. During the coding process used for central controller instructions, the CCI object contains the label information and is carried within PCInitiate or PCRpt message.

This document specify two new CCI object-types for VLAN-based traffic forwarding in the Native IP network and are said to be mandatory in a PCEP message when the object must be included and are considered to be valid. In addition, this document extends the PCEP message to handle the VLAN-based traffic forwarding path in the native IP network with the new CCI object.

5.1. The PCInitiate message

The PCInitiate message[RFC8281] extended in[RFC9050] can be used to download or remove labels by using the CCI Object.

Based on the extended PCInitiate message and PCRpt described in [I-D.ietf-pce-pcep-extension-native-ip], the (BGP Peer Info (BPI) Object and the Peer Prefix Association (PPA) Object are used to establish multi BGP sessions and advertise route prefixes among different BGP sessions before setting up a VLAN-based traffic forwarding path.

This document extends the PCInitiate message as shown below:

```
<PCInitiate Message> ::= <Common Header>
                              <PCE-initiated-lsp-list>
   Where:
      <Common Header> is defined in [RFC5440]

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

      <PCE-initiated-lsp-request> ::=
                            (<PCE-initiated-lsp-instantiation>|
                             <PCE-initiated-lsp-deletion>|
                             <PCE-initiated-lsp-central-control>)

      <PCE-initiated-lsp-central-control> ::= <SRP>
                                              <LSP>
                                              <cci-list>|
                                              ((<BPI>|<PPA>)
                                              <new-CCI>)

      <cci-list> ::=  <new-CCI>
                      [<cci-list>]
```

Where:

<cci-list> is as per [RFC9050].
<PCE-initiated-lsp-instantiation> and <PCE-initiated-lsp-deletion> are as per [RFC8281].
<BPI> and <PPA> are as per [draft-ietf-pce-pcep-extension-native-ip-09]

When PCInitiate message is used to create VLAN-based forwarding instructions, the SRP, LSP and CCI objects MUST be present. The error handling for missing SRP, LSP or CCI object is as per [RFC9050]. Further only one of BPI, PPA or one type of CCI objects MUST be present. If none of them are present, the receiving PCE MUST send a PCErr message with Error-type=6 (Mandatory Object missing) and Error-value=TBD4 (VLAN-based forwarding object missing). If there are more than one of BPI, PPA or more than one type of CCI objects, the receiving PCC MUST send a PCErr message with Error-type=19(Invalid Operation) and Error-value=TBD5(Only one of BPI, PPA or one type of the CCI objects for VLAN can be included in this message).

5.2. The PCRpt message

The PCRpt message is used to report the state and confirm the VLAN info that was allocated by the PCE, to be used during the state synchronization phase or as acknowledgement to PCInitiate message.

The format of the PCRpt message is as follows:

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

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

         <state-report> ::= (<lsp-state-report>|
                            <central-control-report>)

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

         <central-control-report> ::= [<SRP>]
                                      <LSP>
                                      <cci-list>|
                                      ((<BPI>|<PPA>)
                                      (<new-CCI>)


        Where:
         <path> is as per [RFC8231] and the LSP and SRP object are
         also defined in [RFC8231].
         <BPI> and <PPA> are as per
         [draft-ietf-pce-pcep-extension-native-ip-09]
```

The error handling for missing LSP or CCI object is as per [RFC9050]. Further only one of BPI, PPA or one type of CCI objects MUST be present. If none of them are present, the receiving PCE MUST send a PCErr message with Error-type=6 (Mandatory Object missing) and Error-value=TBD4 (VLAN-based forwarding object missing). If there are more than one of BPI, PPA or more than one type of CCI objects, the receiving PCC MUST send a PCErr message with Error-type=19(Invalid Operation) and Error-value=TBD5(Only one of BPI, PPA or one type of the CCI objects for VLAN can be included in this message).

6. VSP Operations

Based on [RFC8281] and [RFC9050], in order to set up a PCE-initiated VSP based on the PCECC mechanism, a PCE needs to send a PCInitiate message with the PST set to TBD1 in SRP for the PCECC to the ingress PCC.

The VLAN-forwarding instructions from the PCECC needs to be sent after the initial PCInitiate and PCRpt message exchange with the ingress PCC. On receipt of a PCInitiate message for the PCECC VSP,

the PCC responds with a PCRpt message with the status set to 'Going-up', carrying the assigned PLSP-ID and set the D(Delegate) flag and C(Create) flag(see Figure 1).

After that, the PCE needs to send a PCInitiate message to each node along the path to download the VLAN instructions. The new CCI for the VLAN operations in PCEP are sent via the PCInitiate message by defining a new PCEP object for CCI operations. The fields in the LSP-IDENTIFIERS TLV are described for the RSVP-signaled LSPs but are applicable to the PCECC VSP as well. So the LSP object is included in the PCInitiate message can still be used to identify the PCECC VSP for this instruction and the process is the same.

When the PCE receives this PCRpt message with the PLSP-ID, it assigns VLANs along the path and sets up the path by sending a PCInitiate message to each node along the path of the VSP, as per the PCECC technique. The ingress PCC would receive one VLAN forwarding CCI Object which contains VLAN on the logical subinterface and the Peer IP address. The transit PCC would receive two VLAN crossing CCI Objects with the O bit set for the out-VLAN on the egress subinterface and the O bit unset for the in-VLAN on the ingress subinterface. Similar with the transit PCC, the egress PCC would receive two VLAN crossing CCI Objects but the out-VLAN on the egress subinterface is set to 0. Once the VLAN operations are completed, the PCE MUST send a PCUpd message to the ingress PCC.

```
                  +-------+                              +-------+
                  |PCC    |                              |  PCE  |
                  |ingress|                              +-------+
          +------|       |                                   |
          | PCC  +-------+                                   |
          | transit| |                                       |
   +------|        | |<--PCInitiate,PLSP-ID=0,PST=TBD1------| PCECC VSP
   |PCC   +--------+ |----PCRpt,PLSP-ID=2,D=1,C=1---------->| Initiate
   |egress  |  |     |         (GOING-UP)                   | PCECC VSP
   +--------+  |     |                                      |
       |       |     |                                      |
       |<-------PCInitiate,VLAN-CROSSING-CC-ID=X1,X2--------| VLAN
       |       | PLSP-ID=2,IN-VLAN=N1,OUT-VLAN=0              | download
       |       |     |                                      | CCI
       |--------PCRpt,VLAN-CROSSING-CC-ID=X1,X2------------>|
       |       |PLSP-ID=2,IN-VLAN=N1,OUT-VLAN=0               |
       |       |     |                                      |
       |       |<---PCInitiate,VLAN-CROSSING-CC-ID=Y1,Y2--->| VLAN
       |       |     |PLSP-ID=2,IN-VLAN=N2,OUT-VLAN=N1        | download
       |       |     |                                      | CCI
       |       |-------PCRpt,VLAN-CROSSING-CC-ID=Y1,Y2----->|
       |       |     |PLSP-ID=2,IN-VLAN=N2,OUT-VLAN=N1        |
       |       |     |                                      |
       |       |     |<--PCInitiate,VLAN-FORWARDING-CC-ID=Z-| VLAN
       |       |     |          PLSP-ID=2,VLAN=N2           | download
       |       |     |                                      | CCI
       |       |     |-----PCRpt,CC-ID=Z,PLSP-ID=2--------->|
       |       |     |          PLSP-ID=2,VLAN=N2           |
       |       |     |                                      |
       |       |     |<---PCUpd,PLSP-ID=2,PST=TBD1,D=1------| PCECC VSP
       |       |     |      (UP)                            | Update
       |       |     |----PCRpt,PLSP-ID=2,D=1,C=1---------->|
       |       |     |      (UP)                            |
```

Figure 1: PCE-Initiated PCECC VSP

In order to delete an LSP based on the PCECC, the PCE sends CCI and SRP object with the R bit set to 1 via a PCInitiate message to each node along the path of the VSP to clean up the label-forwarding instruction.

As per [RFC9050], the PCECC VSP also follows the same make-before-break principles. As shown in the figure 2, new path for VSP triggers the new CCI Distribution process. The PCECC firstly updates the new VLAN instructions and informs each node along the new path through the new VLAN crossing CCI Objects and VLAN forwarding CCI Objects to download the new VSP. The PCUpd message then triggers the traffic switch on the updated path. On receipt of the PCRpt message corresponding to the PCUpd message, the PCE does the cleanup operation for the former VSP,which is the same as the LSP update process.

```
                 +-------+                              +-------+
                 |PCC    |                              |  PCE  |
                 |ingress|                              +-------+
          +------|       |                                  |
          | PCC  +-------+                                  |
          | transit| |                                     |
   +------|        | |                                     |
   |PCC   +--------+ |                                     |
   |egress  |  |     |                                     |
   +--------+  |     |                                     |
       |       |     |                                     |
       |<----- PCInitiate,VLAN-CROSSING-CC-ID=NEW-X1,X2----| New Path
       |       |   PLSP-ID=1,IN-VLAN=NEW-N1,OUT-VLAN=0     | for VSP
       |       |     |                                     | triggers
       |--------PCRpt,VLAN-CROSSING-CC-ID=NEW-X1,X2------->| new CCI
       |       | PLSP-ID=1,IN-VLAN=NEW-N1,OUT-VLAN=0       |
       |       |     |                                     |
       |       |<----------PCInitiate,PLSP-ID=1------------|
       |       |     |VLAN-CROSSING-CC-ID=NEW-Y1,NEW-Y2    | Label
       |       |     |  IN-VLAN=NEW-N2,OUT-VLAN=NEW-N1     | download
       |       |     |                                     | CCI
       |       |--------------PCRpt,PLSP-ID=1------------->|
       |       |     |VLAN-CROSSING-CC-ID=NEW-Y1,NEW-Y2    |
       |       |     |  IN-VLAN=NEW-N2,OUT-VLAN=NEW-N1     |
       |       |     |                                     |
       |       |     |<--------PCInitiate,PLSP-ID=1--------| Label
       |       |     |     VLAN-FORWARDING-CC-ID=NEW-Z     | download
       |       |     |            VLAN=NEW-N2              | CCI
       |       |     |                                     |
       |       |     |----------PCRpt,PLSP-ID=1----------->|
       |       |     |     LAN-FORWARDING-CC-ID=NEW-Z      |
       |       |     |           VLAN=NEW-N2               |
       |       |     |                                     |
       |       |     |<---PCUpd,PLSP-ID=1,PST=TBD1,D=1-----| PCECC
       |       |     |    (SRP=S)                          | VSP Update
       |       |     |                                     |
       |       |     |---PCRpt,PLSP-ID=1,PST=TBD1,D=1----->| Trigger
```

```
        |        |      |        (SRP=S)                        | Delete
        |        |      |                                       | former CCI
        |        |      |                                       |
        |<--------------PCInitiate, PLSP-ID=1---------------| Label
        |        |      |VLAN-CROSSING-CC-ID=X1,X2,R=1          | cleanup
        |----------------PCRpt,PLSP-ID=1------------------->| CCI
        |        |      |VLAN-CROSSING-CC-ID=X1,X2,R=1          |
        |        |      |                                       |
        |        |<------------PCInitiate,PLSP-ID=1----------| Label
        |        |      |VLAN-CROSSING-CC-ID=Y1,Y2,R=1          | cleanup
        |        |---------------PCRpt,PLSP-ID=1------------>| CCI
        |        |      |VLAN-CROSSING-CC-ID=Y1,Y2,R=1          |
        |        |      |                                       |
        |        |      |<--------PCInitiate,PLSP-ID=1--------| Label
        |        |      |VLAN-FORWARDING-CC-ID=Z,R=1            | cleanup
        |        |      |---------PCRpt,PLSP-ID=1------------>| CCI
        |        |      |VLAN-FORWARDING-CC-ID=Z,R=1            |
```

Figure 2: PCECC VSP Update

7. VLAN-based traffic forwarding Procedures

The target deployment environment of VLAN based traffic forwarding mechanism is for both Native IPv4 and IPv6. In such scenarios, the BGP is used for the prefix distribution among underlying devices(PCCs), no MPLS is involved.

In order to set up the VLAN-based traffic forwarding paths for different applications in native IP network, multiple BGP sessions should be deployed between the ingress PCC and egress PCC at the edge of the network respectively.

Based on the business requirements, the PCE calculates the explicit route and sends the route information to the PCCs through PCInitiate messages. When the PCInitiate message is received, the packet to be guaranteed will be labeled with corresponding VLAN tag, that is done by the ingress PCC. The labeled packet will be further sent to the PCC’s specific subinterface identified by the VLAN tag and then be forwarded. Similarly, after receive of the PCInitiate message, the packet will be re-labeled with new VLAN tag and then be forwarded by the transit PCC and the egress PCC. The mechanism of allocating and managing VLAN ID by PCC is as per [I-D.wang-vlan-based-traffic-forwarding].

The whole procedures mainly focused on the end-to-end traffic assurance for key applications so that it can ensure the adequacy of VLAN quantity. During the packet forwarding process, the packet can be encapsulated with reserved multicast MAC addresses(e.g. 0180:C200:0014 for ISIS level 1, 0180:C200:0015 for ISIS level 2) thus does not need to be changed hop by hop by each PCC.

7.1. Multiple BGP Session Establishment Procedures

As described in section 4, multiple BGP sessions should be deployed between the ingress device and egress device at the edge of the network respectively in order to carry information of different applications. As per [I-D.ietf-pce-pcep-extension-native-ip], the PCE should send the BPI (BGP Peer Info) Object to the ingress and egress device with the indicated Peer AS and Local/Peer IP address. The Ingress and egress devices will receive multiple BPI objects to establish sessions with different next hop. The specific process is as follows:

```
            +----------------------+
  +---------+-        PCE          + --------+
  |         +----------^-----------+         |
  |          |         |          |          |
  |        +--+       +--+       +--+        |
  |------- +R2+ ------+R3+-------+R4+ --------
  |        +--+       +--+       +--+        |
  |                                          |
  +--+                +--+                +--+
  +R1+----------------+R5+----------------+R6+
  +--+                +--+                +--+
  |                                          |
  |<------------- BGP Session A ------------>|
  |<------------- BGP Session B ------------>|
  |<------------- BGP Session C ------------>|
```

Figure 3: BGP Session Establishment Procedures

7.2. BGP Prefix Advertisement Procedures

The detail procedures for BGP prefix advertisement procedures is introduced in [I-D.ietf-pce-pcep-extension-native-ip], using PCInitiate and PCRpt message pair.

The BGP prefix for different BGP sessions should be sent to the ingress and egress device respectively. The end-to-end traffic for key application can be identified based on these BGP prefix informations and be further assured. As per [I-D.ietf-pce-pcep-extension-native-ip], the PPA(Peer Prefix

Association) object with list of prefix subobjects and the peer address will be sent through the PCInitiate and PCRpt message pair. Through BGP protocol, the ingress device can learn different BGP prefix of the egress device based on the different sessions.

7.3. VLAN mapping info Advertisement Procedures

After the BGP prefix for different BGP session are successfully advertised, information of different applications should be forwarded to different VLAN-based traffic forwarding paths. In order to set up a VLAN-based traffic forwarding path, the PCE should send the VLAN forwarding CCI Object with the VLAN-ID included to the ingress PCC and the VLAN crossing CCI Object to the transit PCC and egress PCC.

7.3.1. VLAN-Based forwarding info Advertisement Procedures

The detail procedures for VLAN-Based forwarding info advertisement contained in the VLAN forwarding CCI Object are shown below, using PCInitiate and PCRpt message pair.

The VLAN forwarding CCI Object should be sent through the PCInitiate and PCRpt message pair. After the PCC receives the CCI object (with the R bit set to 0 in SRP object) in PCInitiate message, the PCC's subinterface will set up the specific VLAN based on the VLAN forwarding CCI object, source and destination BGP prefix learnt before. When the ingress PCC receives a packet, based on the source and destination IP, the packet that needs to be guaranteed will be matched and then be labeled with corresponding VLAN tag. After that, The labeled packet will be further forwarded to the specific subinterface as per [I-D.wang-vlan-based-traffic-forwarding].

When PCC receives the VLAN forwarding CCI Object with the R bit set to 1 in SRP object in PCInitiate message, the PCC should withdraw the VLAN-Based forwarding info advertisement to the peer that indicated by this object.

On receipt of a PCInitiate message for the PCECC VSP, the PCC should report the result via the PCRpt messages, with the corresponding SRP and CCI object included.

```
                +----------------------+
     +---------+         PCE          + --------+
     |         +----------^-----------+         |
     |          |          |          |          |
    M1&M1-R     |          |          |          |
     |          |          |          |          |
     |          |          |          |          |
     |         +--+       +--+       +--+        |
     |------- +R2+ ------+R3+-------+R4+ --------
     |         +--+       +--+       +--+        |
     |                                           |
     +--+                +--+                +--+
     +R1+----------------+R5+----------------+R6+
     +--+                +--+                +--+
```

Figure 4: VLAN-Based forwarding info Advertisement Procedures for Ingress PCC

The message number, message peers, message types and message key parameters in the above figures are shown in the table below:

Table 1: Message Information

No.	Peers	Type	Message Key Parameters
M1 M1-R	PCE/R1	PCInitiate PCRpt	CC-ID=X1(Symbolic Path Name=Class A) VLAN Forwarding CCI Object (Peer_IP=R6_A,Interface_Address=INF1, VLAN_ID=VLAN_R1_R2)

7.3.2. VLAN-Based crossing info Advertisement Procedures

The detail procedures for VLAN-Based crossing info advertisement contained in the VLAN crossing CCI Object are shown below, using PCInitiate and PCRpt message pair.

The PCC would receive VLAN crossing CCI Objects with the in-VLAN CCI without the O bit set and the out-VLAN CCI with the O bit set. The in-VLAN tag and an out-VLAN tag in the CCI Objects specifies a new VLAN forwarding path. After the process of VLAN-Based forwarding info advertisement mentioned above, the PCC's subinterface will set up the specific VLAN based on the VLAN crossing CCI Object(with the R bit set to 0 in SRP object) contained in the PCInitiate message. When the transit PCC receives a data packet that has been labeled with VLAN by ingress PCC before, based on matching process of the VLAN tag, the in-VLAN tag of this data packet will be replaced by a new out-VLAN tag of the current transit PCC as per [I-D.wang-vlan-based-traffic-forwarding]. The packet with the new VLAN tag will be further forwarded to the next hop.

For the egress PCC, the out-VLAN tag should be 0 which indicates it is the last hop of the transmission. So the egress PCC will directly remove the in-VLAN tag of the packet and the packet will be forwarded.

When PCC receives the VLAN crossing CCI Object with the R bit set to 1 in SRP object in PCInitiate message, the PCC should withdraw the VLAN-Based crossing info advertisement to the peer that indicated by this object.

On receipt of a PCInitiate message for the PCECC VSP, the PCC should report the result via the PCRpt messages, with the corresponding SRP and CCI object included.

When the out-VLAN tag conflicts with a pre-defined VLAN tag or the PCC can not set up a VLAN forwarding path with the out-VLAN tag, an error (Error-type=TBD6, VLAN-based forwarding failure, Error-value=TBD7, VLAN crossing CCI Object peer info mismatch) should be reported via the PCRpt message.

```
                  +----------------------+
         +---------+         PCE          + --------+
         |         +----------^-----------+         |
         |          |         |          |          |
         |       M1&M1-R    M2&M2-R   M3&M3-R    M4&M4-R
         |          |         |          |          |
         |        +--+       +--+       +--+        |
         |------- +R2+ ------+R3+-------+R4+ -------|
         |        +--+       +--+       +--+        |
         |                                          |
         +--+                +--+                +--+
         +R1+----------------+R5+----------------+R6+
         +--+                +--+                +--+
```

Figure 5: VLAN-Based crossing info Advertisement Procedures for transit PCC and egress PCC

The message number, message peers, message type and message key parameters in the above figures are shown in below table:

Table 2: Message Information

No.	Peers	Type	Message Key Parameters
M1 M1-R	PCE/R2	PCInitiate PCRpt	CC-ID=X1(Symbolic Path Name=Class A) VLAN crossing CCI Object(IN) (O=0,Interface_Address=INF1,IN_VLAN_ID=VLAN_R1_R2) VLAN crossing CCI Object(OUT) (O=1,Interface_Address=INF2,OUT_VLAN_ID=VLAN_R2_R3)
M2 M2-R	PCE/R3	PCInitiate PCRpt	CC-ID=X1(Symbolic Path Name=Class A) VLAN crossing CCI Object(IN) (O=0,Interface_Address=INF1,IN_VLAN_ID=VLAN_R2_R3) VLAN crossing CCI Object(OUT) (O=1,Interface_Address=INF2,OUT_VLAN_ID=VLAN_R3_R4)
M3 M3-R	PCE/R4	PCInitiate PCRpt	CC-ID=X1(Symbolic Path Name=Class A) VLAN crossing CCI Object(IN) (O=0,Interface_Address=INF1,IN_VLAN_ID=VLAN_R3_R4) VLAN crossing CCI Object(OUT) (O=1,Interface_Address=INF2,OUT_VLAN_ID=VLAN_R4_R6)
M4 M4-R	PCE/R6	PCInitiate PCRpt	CC-ID=X1(Symbolic Path Name=Class A) VLAN crossing CCI Object(IN) (O=0,Interface_Address=INF1,IN_VLAN_ID=VLAN_R4_R6) VLAN crossing CCI Object(OUT) (O=1,Interface_Address=INF2,OUT_VLAN_ID=0)

8. New PCEP Objects

The Central Control Instructions (CCI) Object is used by the PCE to specify the forwarding instructions is defined in [RFC9050]. This document defines two other CCI object-types for VLAN-based traffic forwarding. All new PCEP objects are compliant with the PCEP object format defined in [RFC5440].

8.1. VLAN forwarding CCI Object

The VLAN forwarding CCI Object is used to set up the specific VLAN forwarding path including the logical subinterface that will be used for traffic forwarding to the specific hop. Combined with this type of CCI Object and the Peer Prefix Association object(PPA) defined in [I-D.ietf-pce-pcep-extension-native-ip], the ingress PCC will identify the traffic that needs to be protected. This object should only be included and sent to the ingress PCC of the end2end path.

CCI Object-Class is 44.

CCI Object-Type is TBD8 for VLAN forwarding info in the native IP network.

```
 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
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                            CC-ID                              |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|            Reserved1           |            Flags              |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|       VLAN-ID          |             Reserved2                |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                                                               |
//                     Interface Address TLV                    //
|                                                               |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                                                               |
//                     Peer IP Address TLV                      //
|                                                               |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                                                               |
//                        Additional TLVs                       //
|                                                               |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 6: VLAN Forwarding CCI Object

The fields in the CCI object are as follows:

CC-ID: is as described in [RFC9050]. Following fields are defined for CCI Object-Type TBD8.

Reserved1(16 bits): is set to zero while sending, ignored on receipt.

Flags(16 bits): is used to carry any additional information pertaining to the CCI. Currently no flag bits are defined.

VLAN ID(12 bits):the ID of the VLAN forwarding path that the PCC will set up on its logical subinterface in order to transfer the packet to the specific hop.

Reserved2(20 bits): is set to zero while sending, ignored on receipt.

Interface Address TLV [RFC8779] MUST be included in this CCI Object-Type TBD8 to specify the interface which will set up the vlan defined in the VLAN Forwarding CCI Object.

The Peer IP Address TLV [RFC8779]MUST be included in this CCI Object-Type TBD8 to identify the end to end TE path in VLAN-based traffic forwarding network and MUST be unique.

8.2. Address TLVs

[RFC8779] defines IPV4-ADDRESS, IPV6-ADDRESS, and UNNUMBERED-ENDPOINT TLVs for the use of Generalized Endpoint. The same TLVs can also be used in the CCI object to find the Peer address that matches egress PCC and further identify the packet to be guaranteed. If the PCC is not able to resolve the peer information or can not find the corresponding ingress device, it MUST reject the CCI and respond with a PCErr message with Error-Type = TBD6 ("VLAN-based forwarding failure") and Error Value = TBD9 ("Invalid egress PCC information").

8.3. VLAN crossing CCI Object

The VLAN crossing CCI object is defined to control the transmission-path of the packet by VLAN-ID. This new type of CCI Object can be carried within a PCInitiate message sent by the PCE to the transit PCC and the egress PCC in the VLAN-based traffic forwarding scenarios.

CCI Object-Class is 44.

CCI Object-Type is TBD10 for VLAN crossing info in the native IP network.

```
 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
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                            CC-ID                              |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|            Reserved1           |            Flags           |O|
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|     VLAN-ID(in/out)    |             Reserved2                |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                                                               |
//                     Interface Address TLV                    //
|                                                               |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                                                               |
//                        Additional TLVs                       //
|                                                               |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 7: VLAN Crossing CCI Object

CC-ID: is as described in [RFC9050]. Following fields are defined for CCI Object-Type TBD10.

Reserved1(16 bits): is set to zero while sending, ignored on receipt.

Flags(16 bits): is used to carry any additional information pertaining to the CCI. Currently, the following flag bit are defined:

* O bit (out-label) : If the bit is set to '1', it specifies the VLAN is the out-VLAN, and it is mandatory to encode the egress interface information(via Interface Address TLVs in the CCI object). If the bit is not set or set to '0', it specifies the VLAN is the in-VLAN, and it is mandatory to encode the ingress interface information.

VLAN ID(12 bits): The ID of the VLAN switching path. When the O bit is set to 0, the VLAN is the in-VLAN and the ID indicates a VLAN forwarding path which is used to identify the traffic that needs to be protected. When the O bit is set to 1, the VLAN is the out-VLAN and it indicates the ID of the VLAN forwarding path that the PCC will set up on its logical subinterface in order to transfer the packet labled with this VLAN ID to the specific hop. To the transit PCC, the value must not be 0 to indicate it is not the last hop of the VLAN-based traffic forwarding path. To the egress PCC, the value must be 0 to indicate it is the last hop of the VLAN-based traffic forwarding path.

Reserved2(8 bits): is set to zero while sending, ignored on receipt.

Interface Address TLV [RFC8779] MUST be included in this CCI Object-Type TBD8 to specify the interface which will set up the vlan defined in the VLAN Forwarding CCI Object.

9. IANA Considerations

9.1. Path Setup Type Registry

[RFC8408] created a sub-registry within the "Path Computation Element Protocol (PCEP) Numbers" registry called "PCEP Path Setup Types". IANA is requested to allocate a new code point within this registry, as follows:

Value	Description	Reference
TBD1	VLAN-Based Traffic Forwarding Path	This document

9.2. PCECC-CAPABILITY sub-TLV's Flag field

[RFC9050] created a sub- registry within the "Path Computation Element Protocol (PCEP) Numbers" registry to manage the value of the PCECC-CAPABILITY sub- TLV's 32-bits Flag field. IANA is requested to allocate a new bit position within this registry, as follows:

Value	Description	Reference
TBD2(V)	VLAN-Based Forwarding CAPABILITY	This document

9.3. PCEP Object Types

IANA is requested to allocate new registry for the PCEP Object Type:

Object-Class Value	Name	Reference
44	CCI Object-Type TBD8: VLAN forwarding CCI TBD10: VLAN crossing CCI	This document

9.4. PCEP-Error Object

IANA is requested to allocate new error types and error values within the "PCEP-ERROR Object Error Types and Values" sub-registry of the PCEP Numbers registry for the following errors:

Error-Type	Meaning	Error-value	Reference
6	Mandatory Object missing	TBD4:VLAN-based forwarding object missing	This document
10	Reception of an invalid object	TBD3:PCECC VLAN-based-forwarding -CAPABILITY bit is not set	This document
19	Invalid Operation	TBD5: Only one of BPI, PPA or one type of the CCI objects for VLAN can be included in this message	This document
TBD6	VLAN-based forwarding failure	TBD7: VLAN crossing CCI Object peer info mismatch	This document
		TBD9: Invalid egress PCC information	This document

10. Normative References

[I-D.ietf-pce-pcep-extension-for-pce-controller]
Li, Z., Peng, S., Negi, M. S., Zhao, Q., and C. Zhou, "Path Computation Element Communication Protocol (PCEP) Procedures and Extensions for Using the PCE as a Central Controller (PCECC) of LSPs", Work in Progress, Internet-Draft, draft-ietf-pce-pcep-extension-for-pce-controller-14, 5 March 2021, <https://datatracker.ietf.org/doc/html/draft-ietf-pce-pcep-extension-for-pce-controller-14>.

[I-D.ietf-pce-pcep-extension-native-ip]
Wang, A., Khasanov, B., Fang, S., Tan, R., and C. Zhu, "Path Computation Element Communication Protocol (PCEP) Extensions for Native IP Networks", Work in Progress, Internet-Draft, draft-ietf-pce-pcep-extension-native-ip-30, 1 February 2024, <https://datatracker.ietf.org/doc/html/draft-ietf-pce-pcep-extension-native-ip-30>.

[I-D.wang-vlan-based-traffic-forwarding]
Wang, Y., Wang, A., Khasanov, B., Qin, F., Chen, H., and C. Zhu, "Dataplane Operations for VLAN Switching", Work in Progress, Internet-Draft, draft-wang-vlan-based-traffic-forwarding-01, 19 March 2024, <https://datatracker.ietf.org/doc/html/draft-wang-vlan-based-traffic-forwarding-01>.

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <https://www.rfc-editor.org/info/rfc2119>.

[RFC4655] Farrel, A., Vasseur, J.-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>.

[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>.

[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>.

[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>.

[RFC8408] Sivabalan, S., Tantsura, J., Minei, I., Varga, R., and J. Hardwick, "Conveying Path Setup Type in PCE Communication Protocol (PCEP) Messages", RFC 8408, DOI 10.17487/RFC8408, July 2018, <https://www.rfc-editor.org/info/rfc8408>.

[RFC8735] Wang, A., Huang, X., Kou, C., Li, Z., and P. Mi, "Scenarios and Simulation Results of PCE in a Native IP Network", RFC 8735, DOI 10.17487/RFC8735, February 2020, <https://www.rfc-editor.org/info/rfc8735>.

[RFC8779] Margaria, C., Ed., Gonzalez de Dios, O., Ed., and F. Zhang, Ed., "Path Computation Element Communication Protocol (PCEP) Extensions for GMPLS", RFC 8779, DOI 10.17487/RFC8779, July 2020, <https://www.rfc-editor.org/info/rfc8779>.

[RFC9050] Li, Z., Peng, S., Negi, M., Zhao, Q., and C. Zhou, "Path Computation Element Communication Protocol (PCEP) Procedures and Extensions for Using the PCE as a Central Controller (PCECC) of LSPs", RFC 9050, DOI 10.17487/RFC9050, July 2021, <https://www.rfc-editor.org/info/rfc9050>.

Authors' Addresses

Yue Wang
China Telecom
Beiqijia Town, Changping District
Beijing
Beijing, 102209
China
Email: wangy73@chinatelecom.cn

Aijun Wang
China Telecom
Beiqijia Town, Changping District
Beijing
Beijing, 102209
China
Email: wangaj3@chinatelecom.cn

Boris Khasanov
Yandex LLC
Ulitsa Lva Tolstogo 16
Moscow
Email: bhassanov@yandex-team.ru

Fengwei Qin
China Mobile
32 Xuanwumenxi Ave.
Beijing
100032
China
Email: qinfengwei@chinamobile.com

Huaimo Chen
Futurewei
Boston,
United States of America
Email: Huaimo.chen@futurewei.com

Chun Zhu
ZTE Corporation
50 Software Avenue, Yuhua District
Nanjing
Jiangsu, 210012
China
Email: zhu.chun1@zte.com.cn