

Networking Working Group
Internet-Draft
Intended status: Standards Track
Expires: January 10, 2022

R. Chen
Zh. Zhang
ZTE Corporation
H. Chen
S. Dhanaraj
Futurewei
F. Qin
China Mobile
A. Wang
China Telecom
July 9, 2021

PCEP Extensions for BIER-TE
draft-chen-pce-bier-09

Abstract

Bit Index Explicit Replication (BIER)-TE shares architecture and packet formats with BIER as described in [RFC8279]. BIER-TE forwards and replicates packets based on a BitString in the packet header, but every BitPosition of the BitString of a BIER-TE packet indicates one or more adjacencies as described in [I-D.ietf-bier-te-arch]. BIER-TE Path can be derived from a Path Computation Element (PCE).

This document specifies extensions to the Path Computation Element Protocol (PCEP) that allow a PCE to compute and initiate the path for the BIER-TE.

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 January 10, 2022.

Copyright Notice

Copyright (c) 2021 IETF Trust and the persons identified as the document authors. All rights reserved.

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

Table of Contents

1. Introduction	3
2. Conventions used in this document	3
3. Overview of PCEP Operation in BIER Networks	3
4. Object Formats	3
4.1. The OPEN Object	4
4.1.1. The BIER-TE PCE Capability sub-TLV	4
4.2. The RP/SRP Object	5
4.3. END-POINTS object	5
4.4. Objective Functions	5
4.5. ERO Object	5
4.5.1. BIER-TE-ERO Subobject	5
4.6. RRO Object	7
5. Procedures	7
5.1. Exchanging the BIER-TE Capability	7
5.2. BIER-TE-ERO Processing	8
5.3. BIER-TE-RRO Processing	8
6. IANA Considerations	8
6.1. PCEP Objects	8
6.1.1. BIER-TE-PCE-CAPABILITY Sub-TLV Type Indicators	9
6.1.2. New Path Setup Type	9
6.1.3. Objective Functions	9
6.1.4. BIER-TE-ERO and RRO Subobjects	9
6.1.5. PCEP-Error Objects and Types	10
7. Security Considerations	10
8. Acknowledgements	10
9. Normative references	10
Authors' Addresses	12

1. Introduction

Bit Index Explicit Replication (BIER)-TE shares architecture and packet formats with BIER as described in [RFC8279]. BIER-TE forwards and replicates packets based on a BitString in the packet header, but every BitPosition of the BitString of a BIER-TE packet indicates one or more adjacencies as described in [I-D.ietf-bier-te-arch]. BIER-TE Path can be derived from a Path Computation Element (PCE).

[RFC8231] specifies a set of extensions to PCEP that allow a PCE to compute and recommend network paths in compliance with [RFC4657] and defines objects and TLVs for MPLS-TE LSPs.

This document uses a PCE for computing one or more BIER-TE paths taking into account various constraints and objective functions.

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. Overview of PCEP Operation in BIER Networks

BIER-TE forwards and replicates packets based on a BitString in the packet header, and every BitPosition of the BitString of a BIER-TE packet indicates one or more adjacencies as described in [I-D.ietf-bier-te-arch]. In a PCEP session, An ERO object specified in [RFC5440] can be extended to carry a BIER-TE path consists of one or more BIER-TE-ERO subobject(s). BIER-TE computed by a PCE can be represented in the following forms:

- o An ordered set of adjacencies BitString(s) in which each bit represents that the adjacencies to which the BFR should replicate packets to in the domain.

In this document, we define a set of PCEP protocol extensions, including a new PCEP capability, a new Path Setup Type (PST), reuse BIER END-POINT Object, a new Objective Functions subobjects, a new ERO subobjects, a new RRO subobjects, a new PCEP error codes and procedures.

4. Object Formats

4.1. The OPEN Object

4.1.1. The BIER-TE PCE Capability sub-TLV

[RFC8408] defines the PATH-SETUP-TYPE-CAPABILITY TLV for use in the OPEN object. The PATH-SETUP-TYPE-CAPABILITY TLV contains an optional list of sub-TLVs which are intended to convey parameters that are associated with the path setup types supported by a PCEP speaker.

This document defines a new Path Setup Type (PST) for BIER-TE as follows:

- o PST = TBD2: Path is setup using BIER-TE technique.

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.

This document also defines the BIER-TE-PCE-CAPABILITY sub-TLV. PCEP speakers use this sub-TLV to exchange BIER capability. If a PCEP speaker includes PST=TBD2 in the PST List of the PATH-SETUP-TYPE-CAPABILITY TLV then it MUST also include the BIER-TE-PCE-CAPABILITY sub-TLV inside the PATH-SETUP-TYPE-CAPABILITY TLV.

The format of the BIER-TE-PCE-CAPABILITY sub-TLV is shown in the following figure:

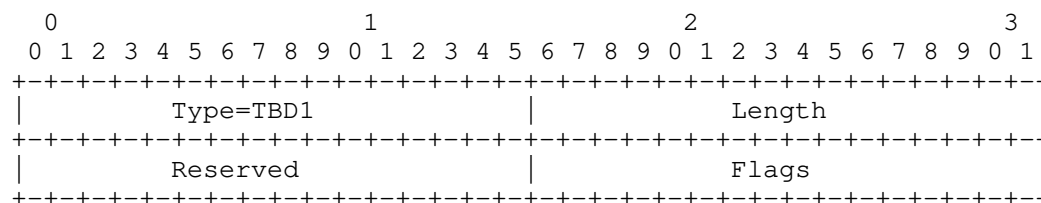


Figure 1 BIER-TE-PCE-CAPABILITY sub-TLV format

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

Length: 4 bytes.

The "Reserved" (2 octet) and "Flags" (2 octet) fields are currently unused, and MUST be set to zero on transmission and ignored on reception.

4.2. The RP/SRP Object

In order to setup an BIER-TE, a new PATH-SETUP-TYPE TLV MUST be contained in RP/SRP object. This document defines a new Path Setup Type (PST=TBD2) for BIER-TE.

4.3. END-POINTS object

The END-POINTS object which is defined in [RFC8306] is used in a PCReq message to specify the BIER information of the path for which a path computation is requested. To represent the end points for a BIER path efficiently, we reuse the P2MP END-POINTS object body for IPv4 (Object-Type 3) and END-POINTS object body for IPv6 (Object-Type 4) which is defined in [RFC8306].

4.4. Objective Functions

[RFC5541] defines a mechanism to specify an objective function (OF) that is used by a PCE when it computes a path. For a BIER-TE path, a new OF is defined.

Objective Function Code: TBD3

Name: Minimum Bit Sets (MBS)

Description: Find a path represented by BitPositions that has the minimum number of bit sets.

4.5. ERO Object

BIER-TE consists of one or more adjacencies BitStrings where every BitPosition of the BitString indicates one or more adjacencies, as described in ([RFC8279]).

The ERO object specified in [RFC5440] is used to encode the path of a TE LSP through the network. The ERO is carried within a PCRep message to provide the computed TE LSP if the path computation was successful. In order to carry BIER-TE explicit paths, this document defines a new ERO subobjects referred to as "BIER-TE-ERO subobjects" whose formats are specified in the following section. An BIER-TE-ERO subobjects carrying a adjacencies BitStrings consists of one or more BIER-TE-ERO subobject(s).

4.5.1. BIER-TE-ERO Subobject

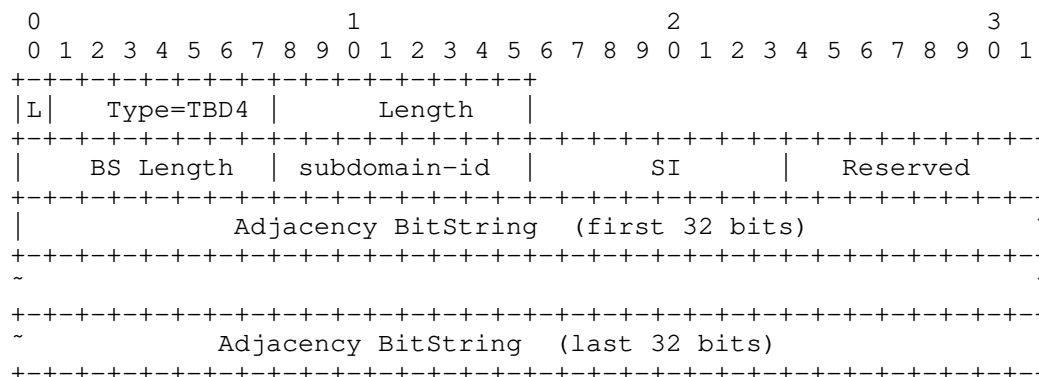


Figure 3

The 'L' Flag: Indicates whether the subobject represents a loose-hop in the LSP[RFC3209]. If the bit is not set, the subobject represents a strict hop in the explicit route.

Type: TBD4

Length: 1 octet ([RFC3209]). Contains the total length of the subobject in octets. The Length MUST be at least 8, and MUST be a multiple of 4.

BS Length: A 1 octet field encodes the length in bits of the BitString as per [RFC8296], the maximum length of the BitString is 5, it indicates the length of BitString is 1024. It is used to refer to the number of bits in the BitString.

subdomain-id: Unique value identifying the BIER subdomain. 1 octet.

SI: Set Identifier (Section 1 of [RFC8279] used in the encapsulation for this BIER subdomain for this BitString length, 1 octet.

The "Reserved" (1 octets) fields are currently unused, and MUST be set to zero on transmission and ignored on reception.

Adjacency BitString: a variable length field encoding the Adjacency BitString where every BitPosition of the BitString indicates one or more adjacencies. the length of this field is according the BS length. The minimum value of this field is 64 bits, and the maximum value of this field is 1024 bits.

Notice:

The maximum value of BS Length is limited to the 1024 bits, in case the BIER-TE-ERO Subobject is too long.

4.6. RRO Object

An RRO contains one or more subobjects called "BIER-TE-RRO subobjects", whose format is shown below:

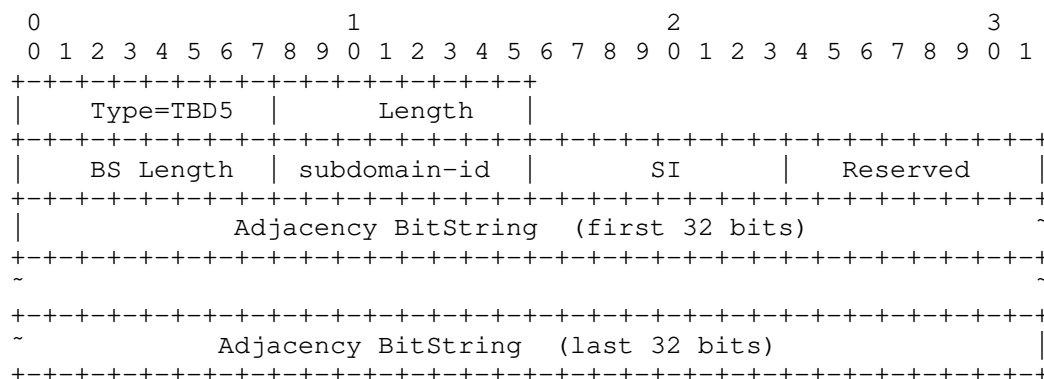


Figure 4

The format of the BIER-TE-RRO subobject is the same as that of the BIER-TE-ERO subobject, but without the L-Flag.

For the integrity of the protocol, we define a new BIER-TE-RRO object, but its actual value is consistent with ERO. The PCC reports an BIER-TE to a PCE by sending a PCRpt message with RRO object.

5. Procedures

5.1. Exchanging the BIER-TE Capability

A PCC indicates that it is capable of supporting the head-end functions for BIER-TE by including the BIER-TE-PCE-CAPABILITY sub-TLV in the Open message that it sends to a PCE. A PCE indicates that it is capable of computing BIER-TE by including the BIET-TE-PCE-CAPABILITY sub-TLV in the Open message that it sends to a PCC.

If a PCEP speaker receives a PATH-SETUP-TYPE-CAPABILITY TLV with a PST list containing PST=TBD2, and supports that path setup type, then it checks for the presence of the SR-PCE-CAPABILITY sub-TLV. If that sub-TLV is absent, then the PCEP speaker MUST send a PCErr message

with Error-Type = 10 ("Reception of an invalid object") and Error-value = TBD6("Missing PCE-BIER-TE-CAPABILITY sub-TLV") and MUST then close the PCEP session. If a PCEP speaker receives a PATH-SETUP-TYPE- CAPABILITY TLV with a BIER-TE-PCE-CAPABILITY sub-TLV, but the PST list does not contain PST=TBD2, then the PCEP speaker MUST ignore the BIER-TE-PCE-CAPABILITY sub-TLV.

5.2. BIER-TE-ERO Processing

If a PCC does not support the BIER-TE PCE Capability and thus cannot recognize the BIER-TE-ERO or BIER-TE-RRO subobjects, The ERO and BIER-TE-ERO subobject processing remains as per [RFC5440].

If a PCC receives an BIER-TE-ERO subobject in which either BitStringLength or Adjacency BitString or SI is absent, it MUST consider the entire BIER-TE-ERO subobject invalid and send a PCErr message with Error-Type = 10 ("Reception of an invalid object"), Error-Value = TBD7 ("BitStringLength is absent ") or Error-Value = TBD8 ("Adjacency BitString is absent") or Error-Value = TBD9 ("SI is absent").

If a PCC receives an BIER-TE-ERO subobject in which BitStringLength values are not chosen from: 64, 128, 256, 512, 1024, as it described in ([RFC8279]). The PCC MUST send a PCErr message with Error-Type =10 ("Reception of an invalid object") and Error-Value = TBD10 ("Invalid BitStringLength").

When a PCEP speaker detects that all subobjects of ERO are not of type TBD4, and if it does not handle such ERO, it MUST send a PCErr message with Error-Type = 10 ("Reception of an invalid object") and Error-Value = TBD11 ("Non-identical ERO subobjects") as per [RFC8664].

5.3. BIER-TE-RRO Processing

The syntax checking rules that apply to the BIER-TE-RRO subobject are identical to those of the BIER-TE-ERO subobject

The actual value of BIER-TE-RRO subobject is consistent with ERO. The PCC reports an BIER-TE to a PCE by sending a PCRpt message with RRO object.

6. IANA Considerations

6.1. PCEP Objects

IANA has made the following Object-Type allocations from the "PCEP Objects" sub-registry.

6.1.1. BIER-TE-PCE-CAPABILITY Sub-TLV Type Indicators

Value	Meaning	Reference
TBD1	BIER-TE-PCE-CAPABILITY	This Document

6.1.2. New Path Setup Type

Value	Meaning	Reference
TBD2	Path is setup using BIER TE technique	This Document

6.1.3. Objective Functions

Value	Meaning	Reference
TBD3	Minimum Bit Sets (MBS)	This Document

6.1.4. BIER-TE-ERO and RRO Subobjects

This document defines a new subobject type for the PCEP explicit route object (ERO) and a new subobject type for the PCEP RRO. The code points for subobject types of these objects are maintained in the RSVP parameters registry, under the EXPLICIT_ROUTE and ROUTE_RECORD objects, respectively.

Object	Subobject	Subobject Type
EXPLICIT_ROUTE	BIER-TE-ERO (PCEP specific)	TBD4
ROUTE_RECORD	BIER-TE-RRO (PCEP specific)	TBD5

6.1.5. PCEP-Error Objects and Types

IANA is requested to allocate code-points in the "PCEP-ERROR Object Error Types and Values" subregistry for the following new error-types and error-values:

Error-Type	Meaning	Error-value
10	Reception of an invalid object	
		TBD6: Missing PCE-BIER-TE-CAPABILITY subobjects
		TBD7: BitStringLength is absent
		TBD8: Adjacency BitString is absent
		TBD9: SI is absent
		TBD10: Invalid BitStringLength
		TBD11: Non-identical ERO subobjects

7. Security Considerations

The security considerations described in [RFC5440], [RFC8231], [RFC8281] and [RFC8408] are applicable to this specification. No additional security measures are required.

8. Acknowledgements

The authors thank Dhruv Dhody, Benchong Xu, Chun Zhu, and Zhaohui Zhang and many others for their suggestions and comments.

9. Normative references

[I-D.ietf-bier-te-arch]
 Eckert, T., Cauchie, G., and M. Menth, "Tree Engineering for Bit Index Explicit Replication (BIER-TE)", draft-ietf-bier-te-arch-09 (work in progress), October 2020.

- [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>>.
- [RFC4657] Ash, J., Ed. and J. Le Roux, Ed., "Path Computation Element (PCE) Communication Protocol Generic Requirements", RFC 4657, DOI 10.17487/RFC4657, September 2006, <<https://www.rfc-editor.org/info/rfc4657>>.
- [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>>.
- [RFC5541] Le Roux, JL., Vasseur, JP., and Y. Lee, "Encoding of Objective Functions in the Path Computation Element Communication Protocol (PCEP)", RFC 5541, DOI 10.17487/RFC5541, June 2009, <<https://www.rfc-editor.org/info/rfc5541>>.
- [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>>.
- [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>>.

- [RFC8306] Zhao, Q., Dhody, D., Ed., Palleti, R., and D. King,
"Extensions to the Path Computation Element Communication
Protocol (PCEP) for Point-to-Multipoint Traffic
Engineering Label Switched Paths", RFC 8306,
DOI 10.17487/RFC8306, November 2017,
<<https://www.rfc-editor.org/info/rfc8306>>.
- [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>>.

Authors' Addresses

Ran Chen
ZTE Corporation

Email: chen.ran@zte.com.cn

Zheng Zhang
ZTE Corporation

Email: zhang.zheng@zte.com.cn

Huaimo Chen
Futurewei

Email: huaimo.chen@futurewei.com

Senthil Dhanaraj
Futurewei

Email: senthil.dhanaraj.ietf@gmail.com

Fengwei Qin
China Mobile

Email: qinfengwei@chinamobile.com

Aijun Wang
China Telecom

Email: wangaj3@chinatelecom.cn

Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: January 12, 2022

H. Chen
M. McBride
Futurewei
G. Mishra
Verizon Inc.
Y. Liu
China Mobile
A. Wang
China Telecom
L. Liu
Fujitsu
X. Liu
Volta Networks
July 11, 2021

PCE for BIER-TE Ingress Protection
draft-chen-pce-bier-te-ingress-protect-00

Abstract

This document describes extensions to Path Computation Element (PCE) communication Protocol (PCEP) for protecting the ingress of a BIER-TE path.

Requirements Language

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

Status of This Memo

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

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at <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 January 12, 2022.

Copyright Notice

Copyright (c) 2021 IETF Trust and the persons identified as the document authors. All rights reserved.

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

Table of Contents

1. Introduction	2
2. Terminologies	3
3. BIER-TE Path Ingress Protection Example	3
4. Behavior around Ingress Failure	4
4.1. Source Detect	5
4.2. Backup Ingress Detect	5
4.3. Both Detect	5
5. Extensions to PCEP	5
5.1. Capability for Ingress Protection	6
5.1.1. Capability for Ingress Protection with Backup Ingress	6
5.1.2. Capability for Ingress Protection with Traffic Source	7
5.2. BIER-TE Path Ingress Protection	8
5.2.1. Extensions for Backup Ingress	8
5.2.2. Extensions for Traffic Source	11
6. IANA Considerations	14
7. Security Considerations	14
8. Acknowledgements	14
9. References	14
9.1. Normative References	14
9.2. Informative References	14
Authors' Addresses	15

1. Introduction

The fast protection of a transit node of a "Bit Index Explicit Replication" (BIER) Traffic Engineering (BIER-TE) path or tunnel is described in [I-D.chen-bier-te-frr]. [RFC8424] presents extensions to RSVP-TE for the fast protection of the ingress node of a traffic engineering (TE) Label Switching Path (LSP). However, these documents do not discuss any protocol extensions for the fast protection of the ingress node of a BIER-TE path or tunnel.

This document fills that gap and specifies protocol extensions to Path Computation Element (PCE) communication Protocol (PCEP) for the fast protection of the ingress node of a BIER-TE path or tunnel. Ingress node and ingress, fast protection and protection as well as BIER-TE path and BIER-TE tunnel will be used exchangeably in the following sections.

2. Terminologies

The following terminologies are used in this document.

PCE: Path Computation Element

PCEP: PCE communication Protocol

PCC: Path Computation Client

BIER: Bit Index Explicit Replication

CE: Customer Edge

PE: Provider Edge

TE: Traffic Engineering

3. BIER-TE Path Ingress Protection Example

Figure 1 shows an example of protecting ingress PE1 of a BIER-TE path, which is from ingress PE1 to egress nodes PE3 and PE4. This primary BIER-TE path is represented by *** in the figure. The ingress of the primary BIER-TE path is called primary ingress.

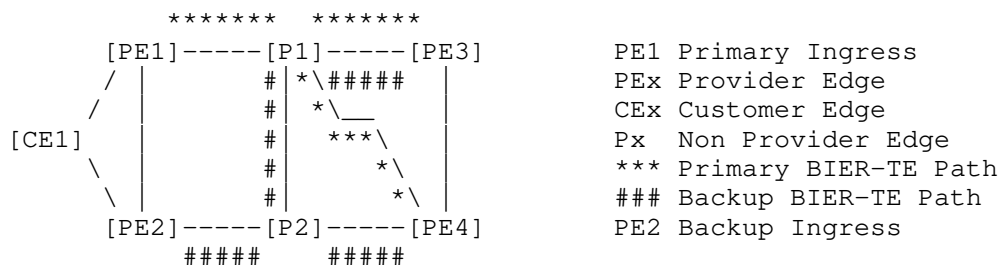


Figure 1: Protecting Ingress PE1 of BIER-TE Path

The backup BIER-TE path is from ingress PE2 to egress nodes PE3 and PE4, which is represented by ### in the figure. The ingress of the backup BIER-TE path is called backup ingress.

In normal operations, CE1 sends the packets with a multicast group and source to ingress PE1, which imports/encapsulates the packets into the BIER-TE path through adding a BIER-TE header. The header contains the BIER-TE path from ingress PE1 to egress nodes PE3 and PE4.

When CE1 detects the failure of ingress PE1 using a failure detection mechanism such as BFD, it switches the traffic to backup ingress PE2, which imports the traffic from CE1 into the backup BIER-TE path. When the traffic is imported into the backup path, it is sent to the egress nodes PE3 and PE4 along the path.

Given the traffic source (e.g., CE1), ingress (e.g., PE1) and egresses (e.g., PE3 and PE4) of the primary BIER-TE path, the PCE computes a backup ingress (e.g., PE2), a backup BIER-TE path from the backup ingress to the egresses, and sends the backup BIER-TE path to the PCC of the backup ingress. It also sends the backup ingress, primary ingress and the traffic description to the PCC of the traffic source (e.g., CE1).

When the PCC of the traffic source receives the backup ingress, primary ingress and traffic description, it sets up the fast detection of the primary ingress failure and the switch over target backup ingress. This setup lets the traffic source node switch the traffic (to be sent to the primary ingress) to the backup ingress when it detects the failure of the primary ingress.

When the PCC of the backup ingress receives the backup BIER-TE path, it adds a forwarding entry into its BIFT. This entry encapsulates the packets from the traffic source in the backup BIER-TE path. This makes the backup ingress send the traffic received from the traffic source to the egress nodes via the backup BIER-TE path.

4. Behavior around Ingress Failure

This section describes the behavior of some nodes connected to the ingress before and after the ingress fails. These nodes are the traffic source (e.g., CE1) and the backup ingress (e.g., PE2). It presents three ways in which these nodes work together to protect the ingress. The first way is called source detect, where the traffic source is responsible for fast detecting the failure of the ingress. The second way is called backup ingress detect, in which the backup ingress is responsible for fast detecting the failure of the ingress. The third way is called both detect, where both the traffic source and the backup ingress are responsible for fast detecting the failure of the ingress.

4.1. Source Detect

In normal operations, i.e., before the failure of the ingress, the traffic source sends the traffic to the ingress of the primary BIER-TE path. The backup ingress (e.g., PE2) is ready to import the traffic from the traffic source into the backup BIER-TE path installed.

When the traffic source detects the failure of the ingress, it switches the traffic to the backup ingress, which delivers the traffic to the egress nodes of the BIER-TE path via the backup BIER-TE path.

4.2. Backup Ingress Detect

The traffic source (e.g., CE1) always sends the traffic to both the ingress (e.g., PE1) of the primary BIER-TE path and the backup ingress (e.g., PE2).

The backup ingress does not import any traffic from the traffic source into the backup BIER-TE path in normal operations. When it detects the failure of the ingress of the primary BIER-TE path, it imports the traffic from the source into the backup BIER-TE path.

For the backup ingress to fast detect the failure of the primary ingress, it SHOULD directly connect to the primary ingress. When a PCE computes a backup ingress and a backup BIER-TE path, it SHOULD consider this.

4.3. Both Detect

In normal operations, i.e., before the failure of the ingress, the traffic source sends the traffic to the ingress of the primary BIER-TE path. When it detects the failure of the ingress, it switches the traffic to the backup ingress.

The backup ingress does not import any traffic from the traffic source into the backup BIER-TE path in normal operations. When it detects the failure of the ingress of the primary BIER-TE path, it imports the traffic from the source into the backup BIER-TE path.

5. Extensions to PCEP

A PCC runs on each of the edge nodes such as PEs and CEs of a network normally. A PCE runs on a server as a controller to communicate with PCCs. The PCE and the PCCs running on backup ingress PEs and traffic source CEs work together to support protection for the ingress of a BIER-TE path.

5.1. Capability for Ingress Protection

5.1.1. Capability for Ingress Protection with Backup Ingress

When a PCE and a PCC running on a backup ingress establish a PCEP session between them, they exchange their capabilities of supporting protection for the ingress node of a BIER-TE path/tunnel.

A new sub-TLV called BIER-TE_INGRESS_PROTECTION_CAPABILITY is defined. It is included in the PATH_SETUP_TYPE_CAPABILITY TLV with PST = TBD1 (suggested value 2 for protecting the ingress of a BIER-TE path/tunnel) in the OPEN object, which is exchanged in Open messages when a PCC and a PCE establish a PCEP session between them. Its format is illustrated below.

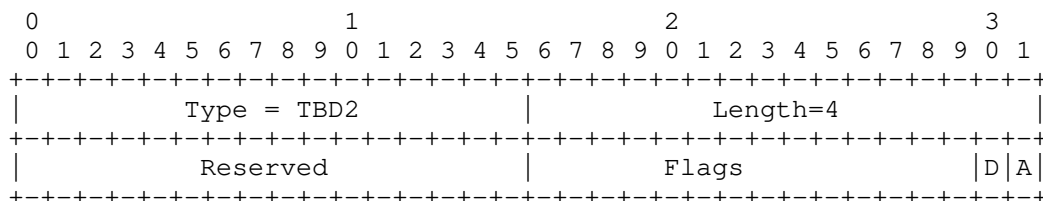


Figure 2: BIER-TE_INGRESS_PROTECTION_CAPABILITY sub-TLV

Type: TBD2 is to be assigned by IANA.

Length: 4.

Reserved: 2 octets. Must be set to zero in transmission and ignored on reception.

Flags: 2 octets. Two flag bits are defined.

- o D flag bit: A PCC sets this flag to 1 to indicate that it is able to detect its adjacent node's failure quickly.
- o A flag bit: A PCE sets this flag to 1 to request a PCC to let the forwarding entry for the backup BIER-TE path/tunnel be Active.

A PCC, which supports ingress protection for a BIER-TE tunnel/path, sends a PCE an Open message containing BIER-TE_INGRESS_PROTECTION_CAPABILITY sub-TLV. This sub-TLV indicates that the PCC is capable of supporting the ingress protection for a BIER-TE tunnel/path.

A PCE, which supports ingress protection for a BIER-TE tunnel/path, sends a PCC an Open message containing BIER-TE_INGRESS_PROTECTION_CAPABILITY sub-TLV. This sub-TLV indicates that the PCE is capable of supporting the ingress protection for a BIER-TE tunnel/path.

If both a PCC and a PCE support BIER-TE_INGRESS_PROTECTION_CAPABILITY, each of the Open messages sent by the PCC and PCE contains PATH-SETUP-TYPE-CAPABILITY TLV with a PST list containing PST=TBD1 and a BIER-TE-INGRESS_PROTECTION_CAPABILITY sub-TLV.

If a PCE receives an Open message without a BIER-TE_INGRESS_PROTECTION_CAPABILITY sub-TLV from a PCC, then the PCE MUST not send the PCC any request for ingress protection of a BIER-TE path/tunnel.

If a PCC receives an Open message without a BIER-TE_INGRESS_PROTECTION_CAPABILITY sub-TLV from a PCE, then the PCC MUST ignore any request for ingress protection of a BIER-TE path/tunnel from the PCE.

If a PCC sets D flag to zero, then the PCE SHOULD send the PCC an Open message with A flag set to one and the fast detection of the failure of the primary ingress MUST be done by the traffic source. When the PCE sends the PCC a message for initiating a backup BIER-TE path, the PCC MUST let the forwarding entry for the backup BIER-TE path be Active.

5.1.2. Capability for Ingress Protection with Traffic Source

When a PCE and a PCC running on a traffic source node establish a PCEP session between them, they exchange their capabilities of supporting protection for the ingress node of a BIER-TE path/tunnel.

The PCECC-CAPABILITY sub-TLV defined in [I-D.ietf-pce-pcep-extension-for-pce-controller] is included in the OPEN object in the PATH-SETUP-TYPE-CAPABILITY TLV, which is exchanged in Open messages when a PCC and a PCE establish a PCEP session between them.

A new flag bit P is defined in the Flags field of the PCECC-CAPABILITY sub-TLV:

- o P flag (for Ingress Protection): if set to 1 by a PCEP speaker, the P flag indicates that the PCEP speaker supports and is willing to handle the PCECC based central controller instructions for ingress protection. The bit MUST be set to 1 by both a PCC and a

PCE for the PCECC ingress protection instruction download/report on a PCEP session.

5.2. BIER-TE Path Ingress Protection

This section specifies the extensions to PCEP for the backup ingress and the traffic source. The extensions let the traffic source

S1: fast detect the failure of the primary ingress and switch the traffic to the backup ingress when the traffic source detects the failure of the primary ingress, or

S2: always send the traffic to both the primary ingress and the backup ingress.

The extensions let the backup ingress

B1: always import the traffic received from the traffic source with possible service ID into the backup BIER-TE path, or

B2: import the traffic with possible service ID into the backup BIER-TE path when the backup ingress detects the failure of the primary ingress.

The following lists the combinations of Si and Bi (i = 1,2) for different ways of failure detects.

Source Detect: S1 and B1.

Backup Ingress Detect: S2 and B2.

Both Detect: S1 and B2.

5.2.1. Extensions for Backup Ingress

For the packets from the traffic source, if the primary ingress (i.e., the ingress of the primary BIER-TE path) encapsulates the packets with a service ID or label into the BIER-TE path, the backup ingress MUST have this service ID or label and encapsulates the packets with the service ID or label into the backup BIER-TE path when the primary ingress fails.

If the backup ingress is requested to detect the failure of the primary ingress, it MUST have the information about the primary ingress such as the address of the primary ingress.

A new TLV called BIER-TE_INGRESS_PROTECTION TLV is defined to transfer the information about the primary ingress and/or the service

ID or label. When a PCE sends the PCC of a backup ingress a PCInitiate message for initiating a backup BIER-TE path/tunnel to protect the primary ingress of a primary BIER-TE path/tunnel, the message contains this TLV in the RP/SRP object. Its format is illustrated below.

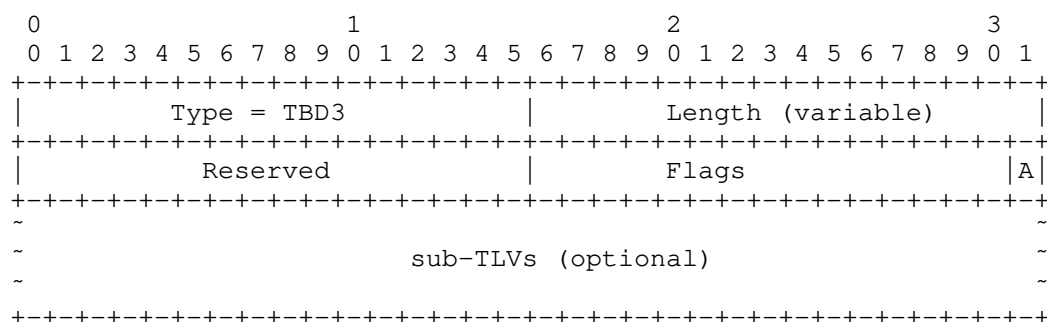


Figure 3: BIER-TE_INGRESS_PROTECTION TLV

Type: TBD3 is to be assigned by IANA.

Length: Variable.

Reserved: 2 octets. Must be set to zero in transmission and ignored on reception.

Flags: 2 octets. One flag bit is defined.

A flag bit: it is set to 1 or 0 by PCE.

- o 1 is to request the backup ingress to let the forwarding entry for the backup BIER-TE path/tunnel be Active always. In this case, the traffic source detects the failure of the primary ingress and switches the traffic to the backup ingress when it detects the failure.
- o 0 is to request the backup ingress to detect the failure of the primary ingress and let the forwarding entry for the backup BIER-TE path/tunnel be Active when the primary ingress fails. In this case, the TLV includes the primary ingress address in a Primary-Ingress sub-TLV. The traffic source can send the traffic to both the primary ingress and the backup ingress. It may switch the traffic to the backup ingress from the primary ingress when it detects the failure of the primary ingress.

Two optional sub-TLVs are defined. One is Service sub-TLV. The other is Primary-Ingress sub-TLV. The Multicast Flow Specification TLV for IPv4 or IPv6, which is defined in [I-D.ietf-pce-pcep-flowspec], is used as a sub-TLV to indicate the traffic to be imported into the backup BIER-TE path.

5.2.1.1. Service sub-TLV

A Service sub-TLV contains a service label such as VPN service label or ID to be added into a packet to be carried by a BIER-TE path/tunnel. It has two formats: one for the service identified by a label and the other for the service identified by a service identifier (ID) of 32 or 128 bits, which are illustrated below.

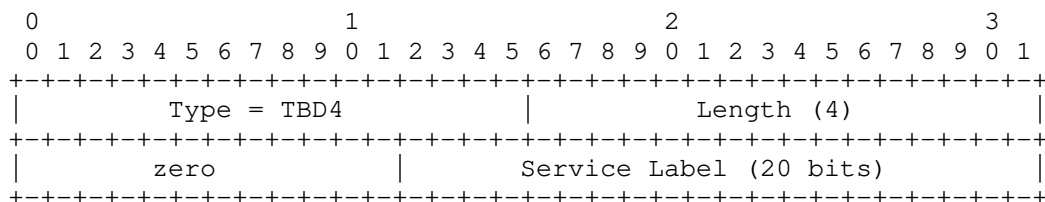


Figure 4: Service Label sub-TLV

Type: TBD4 is to be assigned by IANA.

Length: 4.

Service Label: the least significant 20 bits. It represents a label of 20 bits.

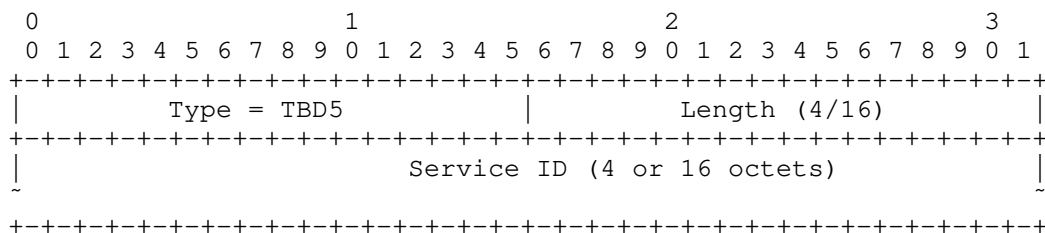


Figure 5: Service ID sub-TLV

Type: TBD5 is to be assigned by IANA.

Length: 4 or 16.

Service ID: 4 or 16 octets. It represents Identifier (ID) of a service in 4 or 16 octets.

5.2.1.2. Primary-Ingress sub-TLV

A Primary-Ingress sub-TLV indicates the IP address of the primary ingress node of a primary BIER-TE path/tunnel. It has two formats: one for primary ingress node IPv4 address and the other for primary ingress node IPv6 address, which are illustrated below.

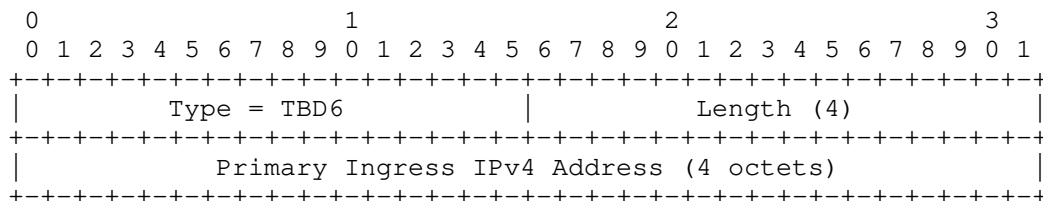


Figure 6: Primary Ingress IPv4 Address sub-TLV

Type: TBD6 is to be assigned by IANA.

Length: 4.

Primary Ingress IPv4 Address: 4 octets. It represents an IPv4 host address of the primary ingress node of a BIER-TE path/tunnel.

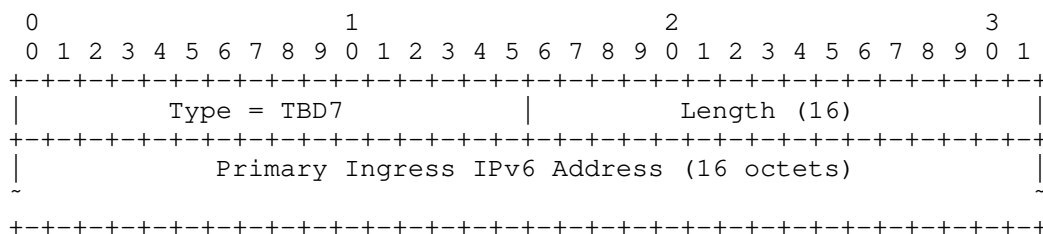


Figure 7: Primary Ingress IPv6 Address sub-TLV

Type: TBD7 is to be assigned by IANA.

Length: 16.

Primary Ingress IPv6 Address: 16 octets. It represents an IPv6 host address of the primary ingress node of a BIER-TE path/tunnel.

5.2.2. Extensions for Traffic Source

If the traffic source is requested to detect the failure of the primary ingress and switch the traffic (to be sent to the primary ingress) to the backup ingress when the primary ingress fails, it MUST have the information about the backup ingress, the primary

ingress and the traffic. This information may be transferred via a CCI object for BIER-TE-INGRESS-PROTECTION to the PCC of the traffic source node from a PCE.

If the traffic source PCC does not accept the request from the PCE or support the extensions, the PCE SHOULD have the information about the behavior of the traffic source configured such as whether it detects the failure of the primary ingress. Based on the information, the PCE instructs the backup ingress accordingly.

The Central Control Instructions (CCI) Object is defined in [I-D.ietf-pce-pcep-extension-for-pce-controller] for a PCE as a controller to send instructions for LSPs to a PCC. This document defines a new object-type (TBDt) for BIER-TE ingress protection based on the CCI object. The body of the object with the new object-type is illustrated below. The object may be in PCRpt, PCUpd, or PCInitiate message.

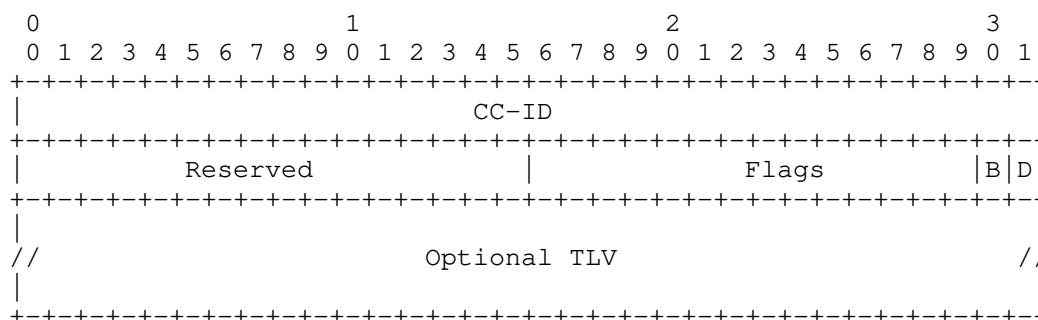


Figure 8: BIER-TE-INGRESS-PROTECTION Object Body

CC-ID: It is the same as described in
[I-D.ietf-pce-pcep-extension-for-pce-controller].

Flags: Two flag bits D and B are defined as follows:

D: D = 1 instructs the PCC of the traffic source to Detect the failure of the primary ingress and switch the traffic to the backup ingress when it detects the failure.

B: B = 1 instructs the PCC of the traffic source to send the traffic to Both the primary ingress and the backup ingress.

Optional TLV: Primary ingress TLV, backup ingress TLV and/or Multicast Flow Specification TLV.

The primary ingress sub-TLV defined above is used as a TLV to contain the information about the primary ingress in the object. The Multicast Flow Specification TLV for IPv4 or IPv6, which is defined in [I-D.ietf-pce-pcep-flowspec], is used to contain the information about the traffic in the object. A new TLV, called backup ingress TLV, is defined to contain the information about the backup ingress in the object.

5.2.2.1. Backup-Ingress TLV

A Backup-Ingress TLV indicates the IP address of the ingress node of a backup BIER-TE path/tunnel. It has two formats: one for backup ingress node IPv4 address and the other for backup ingress node IPv6 address, which are illustrated below. They have the same format as the Primary-Ingress sub-TLVs.

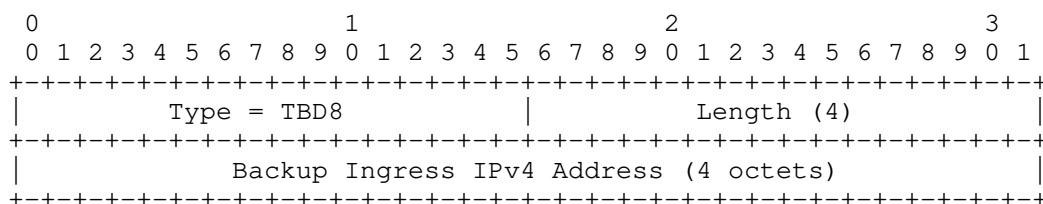


Figure 9: Backup Ingress IPv4 Address TLV

Type: TBD8 is to be assigned by IANA.

Length: 4.

Backup Ingress IPv4 Address: 4 octets. It represents an IPv4 host address of the backup ingress.

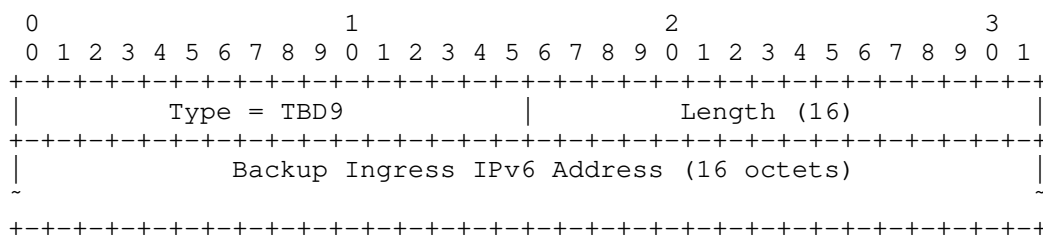


Figure 10: Backup Ingress IPv6 Address TLV

Type: TBD9 is to be assigned by IANA.

Length: 16.

Backup Ingress IPv6 Address: 16 octets. It represents an IPv6 host address of the backup ingress node.

6. IANA Considerations

TBD

7. Security Considerations

TBD

8. Acknowledgements

TBD

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>>.
- [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>>.

9.2. Informative References

- [I-D.chen-bier-te-frr] Chen, H., McBride, M., Liu, Y., Wang, A., Mishra, G. S., Fan, Y., Liu, L., and X. Liu, "BIER-TE Fast ReRoute", draft-chen-bier-te-frr-00 (work in progress), February 2021.

- [I-D.ietf-pce-pcep-extension-for-pce-controller]
Li, Z., Peng, S., Negi, M. S., Zhao, Q., and C. Zhou,
"PCEP Procedures and Protocol Extensions for Using 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-flowspec]
Dhody, D., Farrel, A., and Z. Li, "PCEP Extension for Flow
Specification", draft-ietf-pce-pcep-flowspec-12 (work in
progress), October 2020.
- [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>>.
- [RFC8424] Chen, H., Ed. and R. Torvi, Ed., "Extensions to RSVP-TE
for Label Switched Path (LSP) Ingress Fast Reroute (FRR)
Protection", RFC 8424, DOI 10.17487/RFC8424, August 2018,
<<https://www.rfc-editor.org/info/rfc8424>>.

Authors' Addresses

Huaimo Chen
Futurewei
Boston, MA
USA

Email: Huaimo.chen@futurewei.com

Mike McBride
Futurewei

Email: michael.mcbride@futurewei.com

Gyan S. Mishra
Verizon Inc.
13101 Columbia Pike
Silver Spring MD 20904
USA

Phone: 301 502-1347
Email: gyan.s.mishra@verizon.com

Yisong Liu
China Mobile

Email: liuyisong@chinamobile.com

Aijun Wang
China Telecom
Beiqijia Town, Changping District
Beijing, 102209
China

Email: wangaj3@chinatelecom.cn

Lei Liu
Fujitsu

USA

Email: liulei.kddi@gmail.com

Xufeng Liu
Volta Networks

McLean, VA
USA

Email: xufeng.liu.ietf@gmail.com

Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: 14 July 2022

H. Chen
M. McBride
Futurewei
G. Mishra
Verizon Inc.
Y. Liu
China Mobile
A. Wang
China Telecom
L. Liu
Fujitsu
X. Liu
Volta Networks
10 January 2022

PCE for BIER-TE Ingress Protection
draft-chen-pce-bier-te-ingress-protect-01

Abstract

This document describes extensions to Path Computation Element (PCE) communication Protocol (PCEP) for protecting the ingress of a BIER-TE path.

Requirements Language

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

Status of This Memo

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

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at <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 14 July 2022.

Copyright Notice

Copyright (c) 2022 IETF Trust and the persons identified as the document authors. All rights reserved.

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

Table of Contents

1. Introduction	2
2. Terminologies	3
3. BIER-TE Path Ingress Protection Example	3
4. Behavior around Ingress Failure	4
4.1. Source Detect	5
4.2. Backup Ingress Detect	5
4.3. Both Detect	5
5. Extensions to PCEP	6
5.1. Capability for Ingress Protection	6
5.1.1. Capability for Ingress Protection with Backup Ingress	6
5.1.2. Capability for Ingress Protection with Traffic Source	7
5.2. BIER-TE Path Ingress Protection	8
5.2.1. Extensions for Backup Ingress	9
5.2.2. Extensions for Traffic Source	12
6. IANA Considerations	14
7. Security Considerations	14
8. Acknowledgements	14
9. References	15
9.1. Normative References	15
9.2. Informative References	15
Authors' Addresses	16

1. Introduction

The fast protection of a transit node of a "Bit Index Explicit Replication" (BIER) Traffic Engineering (BIER-TE) path or tunnel is described in [I-D.chen-bier-te-frr]. [RFC8424] presents extensions to RSVP-TE for the fast protection of the ingress node of a traffic engineering (TE) Label Switching Path (LSP). However, these documents do not discuss any protocol extensions for the fast

protection of the ingress node of a BIER-TE path or tunnel.

This document fills that gap and specifies protocol extensions to Path Computation Element (PCE) communication Protocol (PCEP) for the fast protection of the ingress node of a BIER-TE path or tunnel. Ingress node and ingress, fast protection and protection as well as BIER-TE path and BIER-TE tunnel will be used exchangeably in the following sections.

2. Terminologies

The following terminologies are used in this document.

PCE: Path Computation Element

PCEP: PCE communication Protocol

PCC: Path Computation Client

BIER: Bit Index Explicit Replication

CE: Customer Edge

PE: Provider Edge

TE: Traffic Engineering

3. BIER-TE Path Ingress Protection Example

Figure 1 shows an example of protecting ingress PE1 of a BIER-TE path, which is from ingress PE1 to egress nodes PE3 and PE4. This primary BIER-TE path is represented by *** in the figure. The ingress of the primary BIER-TE path is called primary ingress.

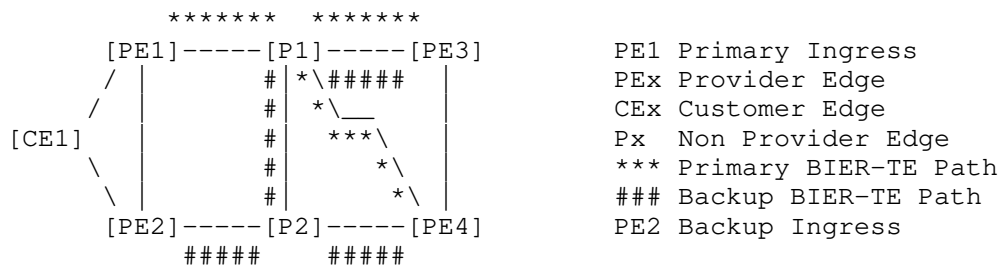


Figure 1: Protecting Ingress PE1 of BIER-TE Path

The backup BIER-TE path is from ingress PE2 to egress nodes PE3 and PE4, which is represented by ### in the figure. The ingress of the backup BIER-TE path is called backup ingress.

In normal operations, CE1 sends the packets with a multicast group and source to ingress PE1, which imports/encapsulates the packets into the BIER-TE path through adding a BIER-TE header. The header contains the BIER-TE path from ingress PE1 to egress nodes PE3 and PE4.

When CE1 detects the failure of ingress PE1 using a failure detection mechanism such as BFD, it switches the traffic to backup ingress PE2, which imports the traffic from CE1 into the backup BIER-TE path. When the traffic is imported into the backup path, it is sent to the egress nodes PE3 and PE4 along the path.

Given the traffic source (e.g., CE1), ingress (e.g., PE1) and egresses (e.g., PE3 and PE4) of the primary BIER-TE path, the PCE computes a backup ingress (e.g., PE2), a backup BIER-TE path from the backup ingress to the egresses, and sends the backup BIER-TE path to the PCC of the backup ingress. It also sends the backup ingress, primary ingress and the traffic description to the PCC of the traffic source (e.g., CE1).

When the PCC of the traffic source receives the backup ingress, primary ingress and traffic description, it sets up the fast detection of the primary ingress failure and the switch over target backup ingress. This setup lets the traffic source node switch the traffic (to be sent to the primary ingress) to the backup ingress when it detects the failure of the primary ingress.

When the PCC of the backup ingress receives the backup BIER-TE path, it adds a forwarding entry into its BIFT. This entry encapsulates the packets from the traffic source in the backup BIER-TE path. This makes the backup ingress send the traffic received from the traffic source to the egress nodes via the backup BIER-TE path.

4. Behavior around Ingress Failure

This section describes the behavior of some nodes connected to the ingress before and after the ingress fails. These nodes are the traffic source (e.g., CE1) and the backup ingress (e.g., PE2). It presents three ways in which these nodes work together to protect the ingress. The first way is called source detect, where the traffic source is responsible for fast detecting the failure of the ingress. The second way is called backup ingress detect, in which the backup ingress is responsible for fast detecting the failure of the ingress. The third way is called both detect, where both the traffic source

and the backup ingress are responsible for fast detecting the failure of the ingress.

4.1. Source Detect

In normal operations, i.e., before the failure of the ingress, the traffic source sends the traffic to the ingress of the primary BIER-TE path. The backup ingress (e.g., PE2) is ready to import the traffic from the traffic source into the backup BIER-TE path installed.

When the traffic source detects the failure of the ingress, it switches the traffic to the backup ingress, which delivers the traffic to the egress nodes of the BIER-TE path via the backup BIER-TE path.

4.2. Backup Ingress Detect

The traffic source (e.g., CE1) always sends the traffic to both the ingress (e.g., PE1) of the primary BIER-TE path and the backup ingress (e.g., PE2).

The backup ingress does not import any traffic from the traffic source into the backup BIER-TE path in normal operations. When it detects the failure of the ingress of the primary BIER-TE path, it imports the traffic from the source into the backup BIER-TE path.

For the backup ingress to fast detect the failure of the primary ingress, it SHOULD directly connect to the primary ingress. When a PCE computes a backup ingress and a backup BIER-TE path, it SHOULD consider this.

4.3. Both Detect

In normal operations, i.e., before the failure of the ingress, the traffic source sends the traffic to the ingress of the primary BIER-TE path. When it detects the failure of the ingress, it switches the traffic to the backup ingress.

The backup ingress does not import any traffic from the traffic source into the backup BIER-TE path in normal operations. When it detects the failure of the ingress of the primary BIER-TE path, it imports the traffic from the source into the backup BIER-TE path.

5. Extensions to PCEP

A PCC runs on each of the edge nodes such as PEs and CEs of a network normally. A PCE runs on a server as a controller to communicate with PCCs. The PCE and the PCCs running on backup ingress PEs and traffic source CEs work together to support protection for the ingress of a BIER-TE path.

5.1. Capability for Ingress Protection

5.1.1. Capability for Ingress Protection with Backup Ingress

When a PCE and a PCC running on a backup ingress establish a PCEP session between them, they exchange their capabilities of supporting protection for the ingress node of a BIER-TE path/tunnel.

A new sub-TLV called BIER-TE_INGRESS_PROTECTION_CAPABILITY is defined. It is included in the PATH_SETUP_TYPE_CAPABILITY TLV with PST = TBD1 (suggested value 2 for protecting the ingress of a BIER-TE path/tunnel) in the OPEN object, which is exchanged in Open messages when a PCC and a PCE establish a PCEP session between them. Its format is illustrated below.

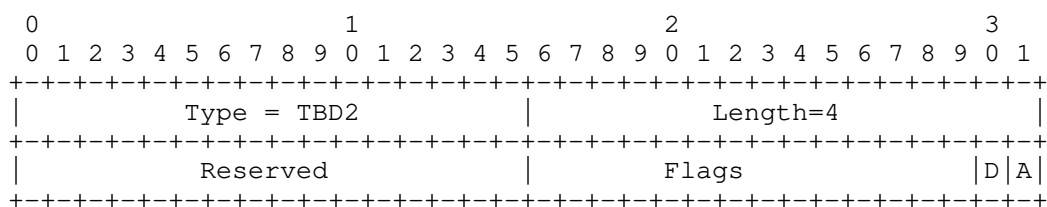


Figure 2: BIER-TE_INGRESS_PROTECTION_CAPABILITY sub-TLV

Type: TBD2 is to be assigned by IANA.

Length: 4.

Reserved: 2 octets. Must be set to zero in transmission and ignored on reception.

Flags: 2 octets. Two flag bits are defined.

- o D flag bit: A PCC sets this flag to 1 to indicate that it is able to detect its adjacent node's failure quickly.
- o A flag bit: A PCE sets this flag to 1 to request a PCC to let

the forwarding entry for the backup BIER-TE path/tunnel be Active.

A PCC, which supports ingress protection for a BIER-TE tunnel/path, sends a PCE an Open message containing BIER-TE_INGRESS_PROTECTION_CAPABILITY sub-TLV. This sub-TLV indicates that the PCC is capable of supporting the ingress protection for a BIER-TE tunnel/path.

A PCE, which supports ingress protection for a BIER-TE tunnel/path, sends a PCC an Open message containing BIER-TE_INGRESS_PROTECTION_CAPABILITY sub-TLV. This sub-TLV indicates that the PCE is capable of supporting the ingress protection for a BIER-TE tunnel/path.

If both a PCC and a PCE support BIER-TE_INGRESS_PROTECTION_CAPABILITY, each of the Open messages sent by the PCC and PCE contains PATH-SETUP-TYPE-CAPABILITY TLV with a PST list containing PST=TBD1 and a BIER-TE-INGRESS_PROTECTION_CAPABILITY sub-TLV.

If a PCE receives an Open message without a BIER-TE_INGRESS_PROTECTION_CAPABILITY sub-TLV from a PCC, then the PCE MUST not send the PCC any request for ingress protection of a BIER-TE path/tunnel.

If a PCC receives an Open message without a BIER-TE_INGRESS_PROTECTION_CAPABILITY sub-TLV from a PCE, then the PCC MUST ignore any request for ingress protection of a BIER-TE path/tunnel from the PCE.

If a PCC sets D flag to zero, then the PCE SHOULD send the PCC an Open message with A flag set to one and the fast detection of the failure of the primary ingress MUST be done by the traffic source. When the PCE sends the PCC a message for initiating a backup BIER-TE path, the PCC MUST let the forwarding entry for the backup BIER-TE path be Active.

5.1.2. Capability for Ingress Protection with Traffic Source

When a PCE and a PCC running on a traffic source node establish a PCEP session between them, they exchange their capabilities of supporting protection for the ingress node of a BIER-TE path/tunnel.

The PCECC-CAPABILITY sub-TLV defined in [I-D.ietf-pce-pcep-extension-for-pce-controller] is included in the OPEN object in the PATH-SETUP-TYPE-CAPABILITY TLV, which is exchanged in Open messages when a PCC and a PCE establish a PCEP session between them.

A new flag bit P is defined in the Flags field of the PCECC-CAPABILITY sub-TLV:

- * P flag (for Ingress Protection): if set to 1 by a PCEP speaker, the P flag indicates that the PCEP speaker supports and is willing to handle the PCECC based central controller instructions for ingress protection. The bit MUST be set to 1 by both a PCC and a PCE for the PCECC ingress protection instruction download/report on a PCEP session.

5.2. BIER-TE Path Ingress Protection

This section specifies the extensions to PCEP for the backup ingress and the traffic source. The extensions let the traffic source

S1: fast detect the failure of the primary ingress and switch the traffic to the backup ingress when the traffic source detects the failure of the primary ingress, or

S2: always send the traffic to both the primary ingress and the backup ingress.

The extensions let the backup ingress

B1: always import the traffic received from the traffic source with possible service ID into the backup BIER-TE path, or

B2: import the traffic with possible service ID into the backup BIER-TE path when the backup ingress detects the failure of the primary ingress.

The following lists the combinations of Si and Bi (i = 1,2) for different ways of failure detects.

Source Detect: S1 and B1.

Backup Ingress Detect: S2 and B2.

Both Detect: S1 and B2.

5.2.1. Extensions for Backup Ingress

For the packets from the traffic source, if the primary ingress (i.e., the ingress of the primary BIER-TE path) encapsulates the packets with a service ID or label into the BIER-TE path, the backup ingress MUST have this service ID or label and encapsulates the packets with the service ID or label into the backup BIER-TE path when the primary ingress fails.

If the backup ingress is requested to detect the failure of the primary ingress, it MUST have the information about the primary ingress such as the address of the primary ingress.

A new TLV called BIER-TE_INGRESS_PROTECTION TLV is defined to transfer the information about the primary ingress and/or the service ID or label. When a PCE sends the PCC of a backup ingress a PCInitiate message for initiating a backup BIER-TE path/tunnel to protect the primary ingress of a primary BIER-TE path/tunnel, the message contains this TLV in the RP/SRP object. Its format is illustrated below.

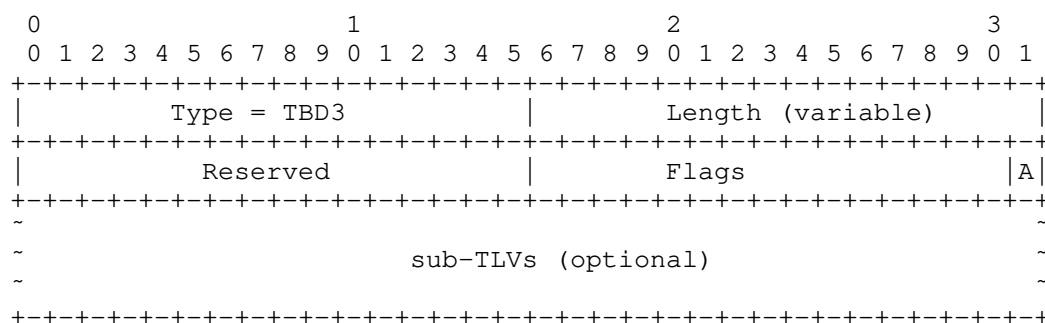


Figure 3: BIER-TE_INGRESS_PROTECTION TLV

Type: TBD3 is to be assigned by IANA.

Length: Variable.

Reserved: 2 octets. Must be set to zero in transmission and ignored on reception.

Flags: 2 octets. One flag bit is defined.

A flag bit: it is set to 1 or 0 by PCE.

o 1 is to request the backup ingress to let the forwarding

entry for the backup BIER-TE path/tunnel be Active always. In this case, the traffic source detects the failure of the primary ingress and switches the traffic to the backup ingress when it detects the failure.

- o 0 is to request the backup ingress to detect the failure of the primary ingress and let the forwarding entry for the backup BIER-TE path/tunnel be Active when the primary ingress fails. In this case, the TLV includes the primary ingress address in a Primary-Ingress sub-TLV. The traffic source can send the traffic to both the primary ingress and the backup ingress. It may switch the traffic to the backup ingress from the primary ingress when it detects the failure of the primary ingress.

Two optional sub-TLVs are defined. One is Service sub-TLV. The other is Primary-Ingress sub-TLV. The Multicast Flow Specification TLV for IPv4 or IPv6, which is defined in [I-D.ietf-pce-pcep-flowspec], is used as a sub-TLV to indicate the traffic to be imported into the backup BIER-TE path.

5.2.1.1. Service sub-TLV

A Service sub-TLV contains a service label such as VPN service label or ID to be added into a packet to be carried by a BIER-TE path/tunnel. It has two formats: one for the service identified by a label and the other for the service identified by a service identifier (ID) of 32 or 128 bits, which are illustrated below.

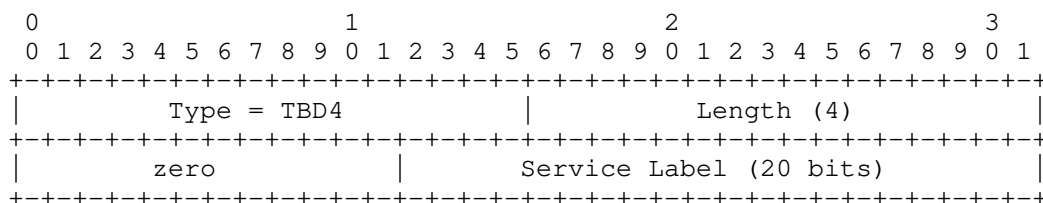


Figure 4: Service Label sub-TLV

Type: TBD4 is to be assigned by IANA.

Length: 4.

Service Label: the least significant 20 bits. It represents a label of 20 bits.

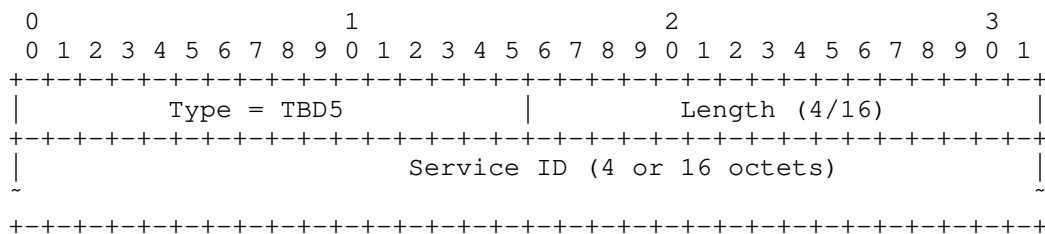


Figure 5: Service ID sub-TLV

Type: TBD5 is to be assigned by IANA.

Length: 4 or 16.

Service ID: 4 or 16 octets. It represents Identifier (ID) of a service in 4 or 16 octets.

5.2.1.2. Primary-Ingress sub-TLV

A Primary-Ingress sub-TLV indicates the IP address of the primary ingress node of a primary BIER-TE path/tunnel. It has two formats: one for primary ingress node IPv4 address and the other for primary ingress node IPv6 address, which are illustrated below.

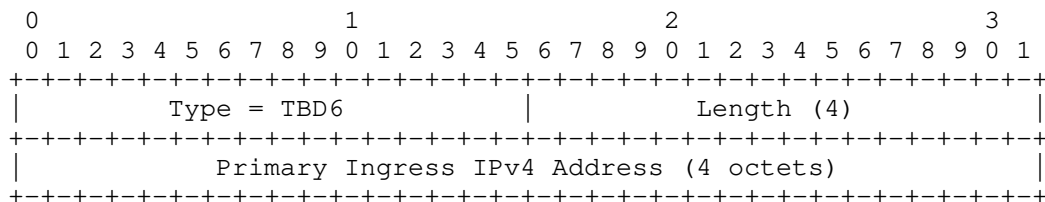


Figure 6: Primary Ingress IPv4 Address sub-TLV

Type: TBD6 is to be assigned by IANA.

Length: 4.

Primary Ingress IPv4 Address: 4 octets. It represents an IPv4 host address of the primary ingress node of a BIER-TE path/tunnel.

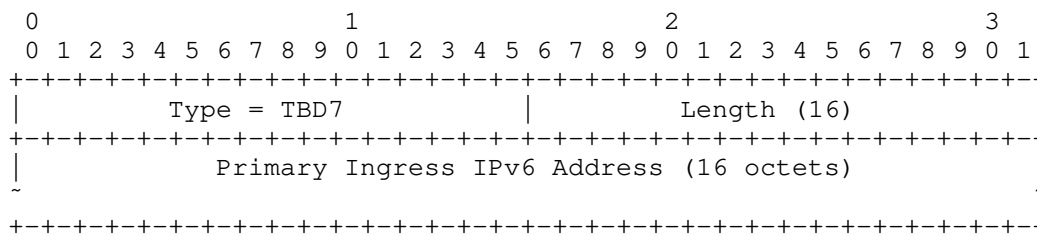


Figure 7: Primary Ingress IPv6 Address sub-TLV

Type: TBD7 is to be assigned by IANA.

Length: 16.

Primary Ingress IPv6 Address: 16 octets. It represents an IPv6 host address of the primary ingress node of a BIER-TE path/tunnel.

5.2.2. Extensions for Traffic Source

If the traffic source is requested to detect the failure of the primary ingress and switch the traffic (to be sent to the primary ingress) to the backup ingress when the primary ingress fails, it MUST have the information about the backup ingress, the primary ingress and the traffic. This information may be transferred via a CCI object for BIER-TE-INGRESS-PROTECTION to the PCC of the traffic source node from a PCE.

If the traffic source PCC does not accept the request from the PCE or support the extensions, the PCE SHOULD have the information about the behavior of the traffic source configured such as whether it detects the failure of the primary ingress. Based on the information, the PCE instructs the backup ingress accordingly.

The Central Control Instructions (CCI) Object is defined in [I-D.ietf-pce-pcep-extension-for-pce-controller] for a PCE as a controller to send instructions for LSPs to a PCC. This document defines a new object-type (TBDt) for BIER-TE ingress protection based on the CCI object. The body of the object with the new object-type is illustrated below. The object may be in PCRpt, PCUpd, or PCInitiate message.

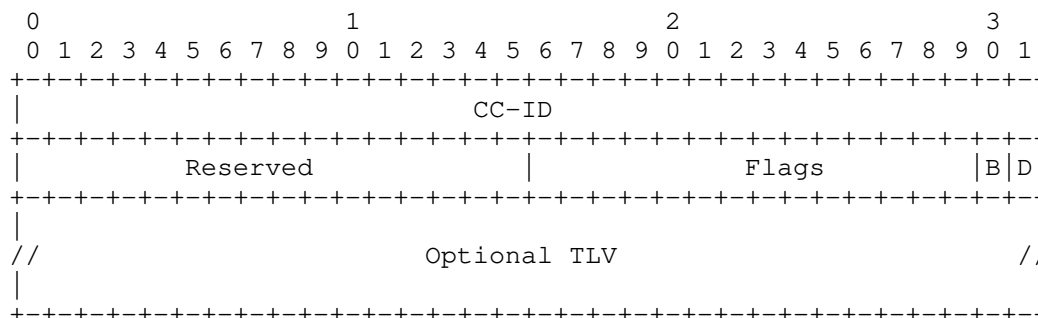


Figure 8: BIER-TE-INGRESS-PROTECTION Object Body

CC-ID: It is the same as described in
[I-D.ietf-pce-pcep-extension-for-pce-controller].

Flags: Two flag bits D and B are defined as follows:

D: D = 1 instructs the PCC of the traffic source to Detect the failure of the primary ingress and switch the traffic to the backup ingress when it detects the failure.

B: B = 1 instructs the PCC of the traffic source to send the traffic to Both the primary ingress and the backup ingress.

Optional TLV: Primary ingress TLV, backup ingress TLV and/or Multicast Flow Specification TLV.

The primary ingress sub-TLV defined above is used as a TLV to contain the information about the primary ingress in the object. The Multicast Flow Specification TLV for IPv4 or IPv6, which is defined in [I-D.ietf-pce-pcep-flowspec], is used to contain the information about the traffic in the object. A new TLV, called backup ingress TLV, is defined to contain the information about the backup ingress in the object.

5.2.2.1. Backup-Ingress TLV

A Backup-Ingress TLV indicates the IP address of the ingress node of a backup BIER-TE path/tunnel. It has two formats: one for backup ingress node IPv4 address and the other for backup ingress node IPv6 address, which are illustrated below. They have the same format as the Primary-Ingress sub-TLVs.

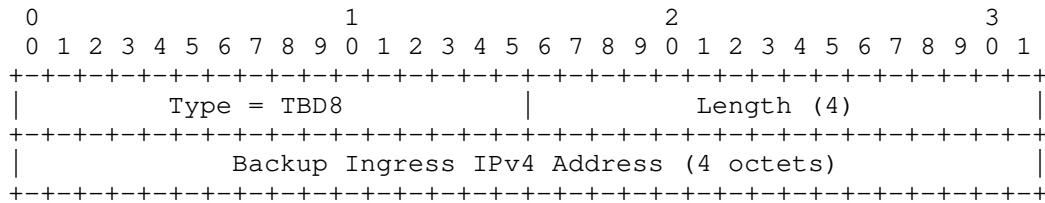


Figure 9: Backup Ingress IPv4 Address TLV

Type: TBD8 is to be assigned by IANA.

Length: 4.

Backup Ingress IPv4 Address: 4 octets. It represents an IPv4 host address of the backup ingress.

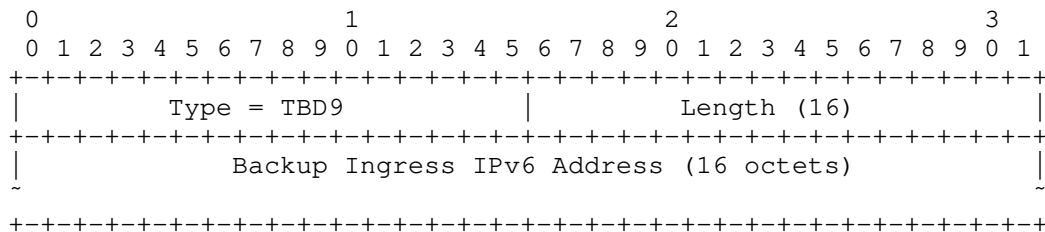


Figure 10: Backup Ingress IPv6 Address TLV

Type: TBD9 is to be assigned by IANA.

Length: 16.

Backup Ingress IPv6 Address: 16 octets. It represents an IPv6 host address of the backup ingress node.

6. IANA Considerations

TBD

7. Security Considerations

TBD

8. Acknowledgements

TBD

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>>.
- [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>>.

9.2. Informative References

- [I-D.chen-bier-te-frr]
Chen, H., McBride, M., Liu, Y., Wang, A., Mishra, G. S., Fan, Y., Liu, L., and X. Liu, "BIER-TE Fast ReRoute", Work in Progress, Internet-Draft, draft-chen-bier-te-frr-01, 23 August 2021, <<https://www.ietf.org/archive/id/draft-chen-bier-te-frr-01.txt>>.
- [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://www.ietf.org/archive/id/draft-ietf-pce-pcep-extension-for-pce-controller-14.txt>>.
- [I-D.ietf-pce-pcep-flowspec]
Dhody, D., Farrel, A., and Z. Li, "PCEP Extension for Flow Specification", Work in Progress, Internet-Draft, draft-ietf-pce-pcep-flowspec-13, 14 October 2021, <<https://www.ietf.org/archive/id/draft-ietf-pce-pcep-flowspec-13.txt>>.

- [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>>.
- [RFC8424] Chen, H., Ed. and R. Torvi, Ed., "Extensions to RSVP-TE for Label Switched Path (LSP) Ingress Fast Reroute (FRR) Protection", RFC 8424, DOI 10.17487/RFC8424, August 2018, <<https://www.rfc-editor.org/info/rfc8424>>.

Authors' Addresses

Huaimo Chen
Futurewei
Boston, MA,
United States of America

Email: Huaimo.chen@futurewei.com

Mike McBride
Futurewei

Email: michael.mcbride@futurewei.com

Gyan S. Mishra
Verizon Inc.
13101 Columbia Pike
Silver Spring, MD 20904
United States of America

Phone: 301 502-1347
Email: gyan.s.mishra@verizon.com

Yisong Liu
China Mobile

Email: liuyisong@chinamobile.com

Aijun Wang
China Telecom
Beiqijia Town, Changping District
Beijing
102209
China

Email: wangaj3@chinatelecom.cn

Lei Liu
Fujitsu
United States of America

Email: liulei.kddi@gmail.com

Xufeng Liu
Volta Networks
McLean, VA
United States of America

Email: xufeng.liu.ietf@gmail.com

PCE
Internet-Draft
Intended status: Standards Track
Expires: January 10, 2022

H. Yuan
UnionPay
T. Zhou
W. Li
G. Fioccola
Y. Wang
Huawei
July 9, 2021

Path Computation Element Communication Protocol (PCEP) Extensions to
Enable IFIT
draft-chen-pce-pcep-ifit-04

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 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 January 10, 2022.

Copyright Notice

Copyright (c) 2021 IETF Trust and the persons identified as the document authors. All rights reserved.

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

Table of Contents

1. Introduction	3
2. PCEP Extensions for IFIT Attributes	4
2.1. IFIT for SR Policies	5
3. IFIT capability advertisement TLV	5
4. IFIT Attributes TLV	7
4.1. IOAM Sub-TLVs	8
4.1.1. IOAM Pre-allocated Trace Option Sub-TLV	9
4.1.2. IOAM Incremental Trace Option Sub-TLV	10
4.1.3. IOAM Directly Export Option Sub-TLV	10
4.1.4. IOAM Edge-to-Edge Option Sub-TLV	11
4.2. Enhanced Alternate Marking Sub-TLV	12
5. PCEP Messages	13
5.1. The PCInitiate Message	13
5.2. The PCUpd Message	14
5.3. The PCRpt Message	14
6. Example of application to SR Policy	14
7. IANA Considerations	15
8. Security Considerations	17
9. Contributors	18
10. Acknowledgements	18
11. References	18
11.1. Normative References	18
11.2. Informative References	20
Appendix A.	21
Authors' Addresses	21

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 attributes, 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.

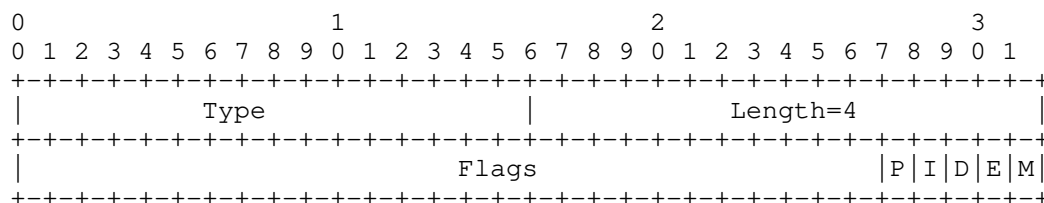


Fig. 1 IFIT-CAPABILITY TLV Format

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:

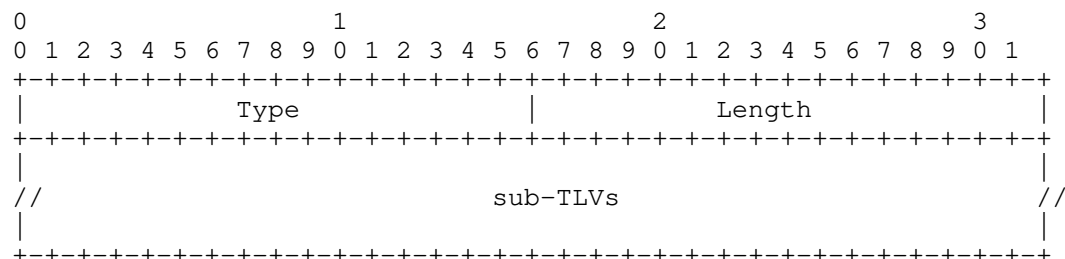


Fig. 2 IFIT-ATTRIBUTES TLV Format

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:

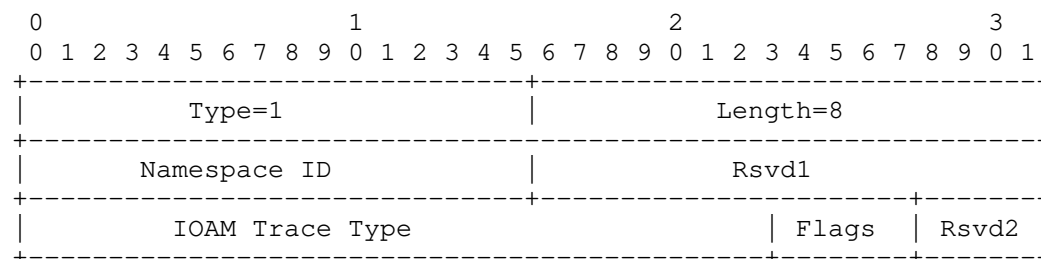


Fig. 4 IOAM Pre-allocated Trace Option Sub-TLV

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:

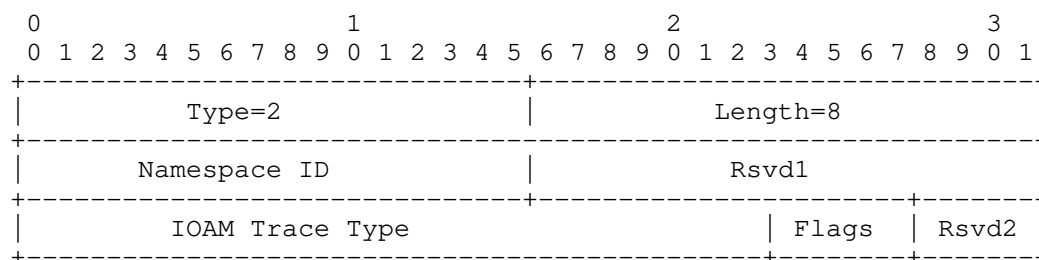


Fig. 5 IOAM Incremental Trace Option Sub-TLV

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:

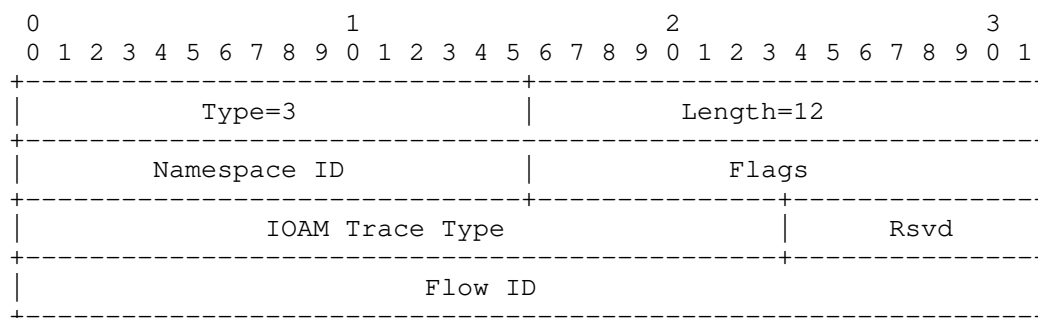


Fig. 6 IOAM Directly Export Option Sub-TLV

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:

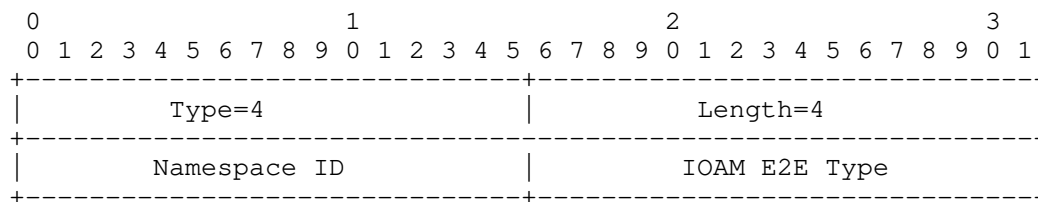


Fig. 7 IOAM Edge-to-Edge Option Sub-TLV

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:

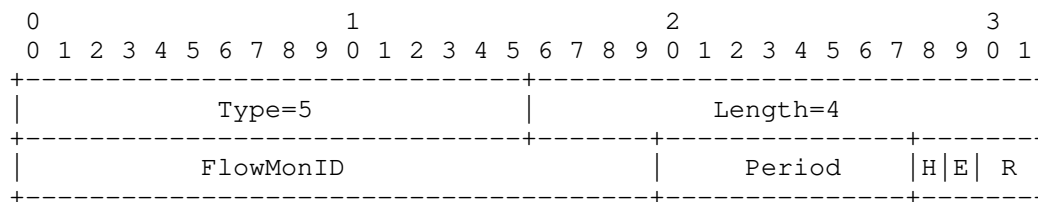


Fig. 8 Enhanced Alternate Marking Sub-TLV

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.

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

E: A flag indicating that the measurement is end to end.

R: A 2-bit field reserved for further usage. It MUST be zero 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. 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	This document
TBD2	IFIT-ATTRIBUTES	This document

This document specifies the IFIT-CAPABILITY TLV 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
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

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

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-04 (work in progress), March 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-12 (work in progress), February 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-03 (work in progress), February 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 Flags", draft-ietf-ippm-ioam-flags-04
(work in progress), February 2021.
- [I-D.ietf-ippm-ioam-ipv6-options]
Bhandari, S., Brockners, F., Pignataro, C., Gredler, H.,
Leddy, J., Youell, S., Mizrahi, T., Kfir, A., Gafni, B.,
Lapukhov, P., Spiegel, M., Krishnan, S., Asati, R., and M.
Smith, "In-situ OAM IPv6 Options", draft-ietf-ippm-ioam-
ipv6-options-05 (work in progress), February 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., Negl, 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-ipv6-09 (work in progress), May 2021.
- [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-04 (work in progress), March 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-11 (work in progress), April 2021.

[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.

Appendix A.

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: wangyalil1@huawei.com

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 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 August 8, 2022.

Copyright Notice

Copyright (c) 2022 IETF Trust and the persons identified as the document authors. All rights reserved.

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

Table of Contents

1. Introduction	3
2. PCEP Extensions for IFIT Attributes	4
2.1. IFIT for SR Policies	5
3. IFIT capability advertisement TLV	5
4. IFIT Attributes TLV	7
4.1. IOAM Sub-TLVs	8
4.1.1. IOAM Pre-allocated Trace Option Sub-TLV	9
4.1.2. IOAM Incremental Trace Option Sub-TLV	10
4.1.3. IOAM Directly Export Option Sub-TLV	10
4.1.4. IOAM Edge-to-Edge Option Sub-TLV	11
4.2. Enhanced Alternate Marking Sub-TLV	12
5. PCEP Messages	13
5.1. The PCInitiate Message	13
5.2. The PCUpd Message	14
5.3. The PCRpt Message	14
6. Example of application to SR Policy	14
7. IANA Considerations	15
7.1. PCEP TLV Type Indicators	15
7.2. IFIT-CAPABILITY TLV Flags field	16
7.3. IFIT-ATTRIBUTES Sub-TLV	16
7.4. Enhanced Alternate Marking Sub-TLV Flags field	17
7.5. PCEP Error Codes	18
8. Security Considerations	18
9. Contributors	19
10. Acknowledgements	19
11. References	19
11.1. Normative References	19
11.2. Informative References	21
Authors' Addresses	22

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 attributes, 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.

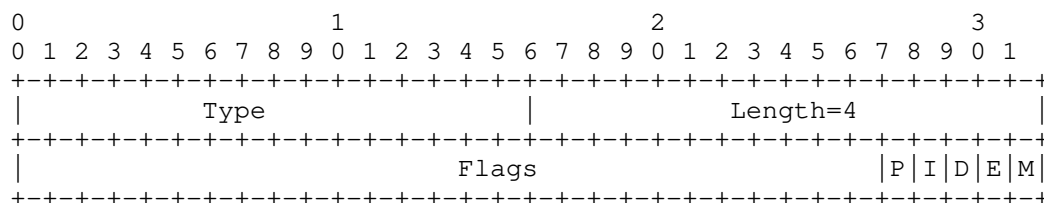


Fig. 1 IFIT-CAPABILITY TLV Format

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:

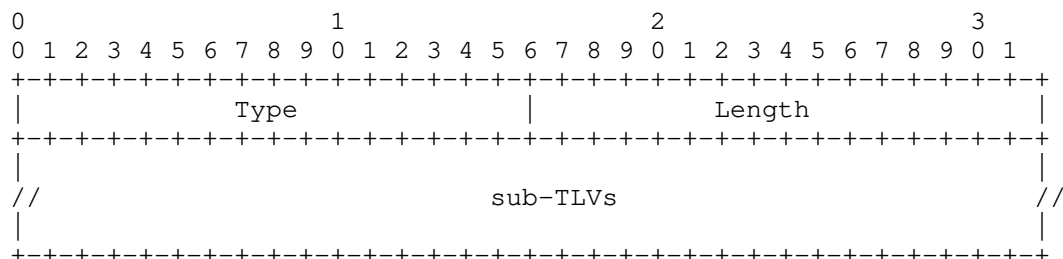


Fig. 2 IFIT-ATTRIBUTES TLV Format

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:

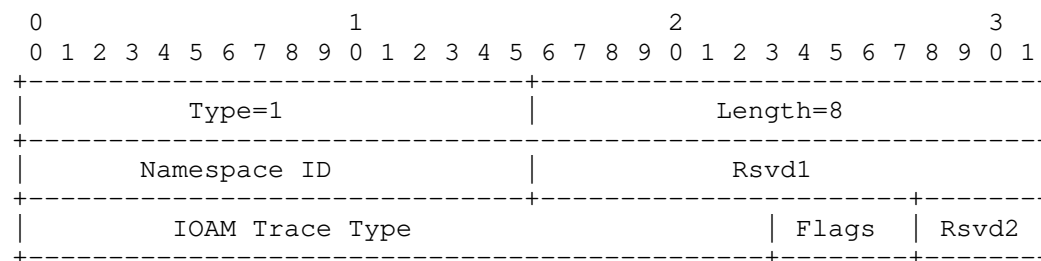


Fig. 4 IOAM Pre-allocated Trace Option Sub-TLV

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:

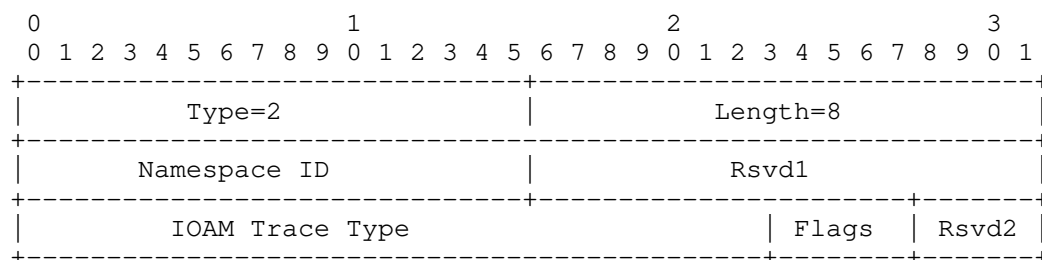


Fig. 5 IOAM Incremental Trace Option Sub-TLV

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:

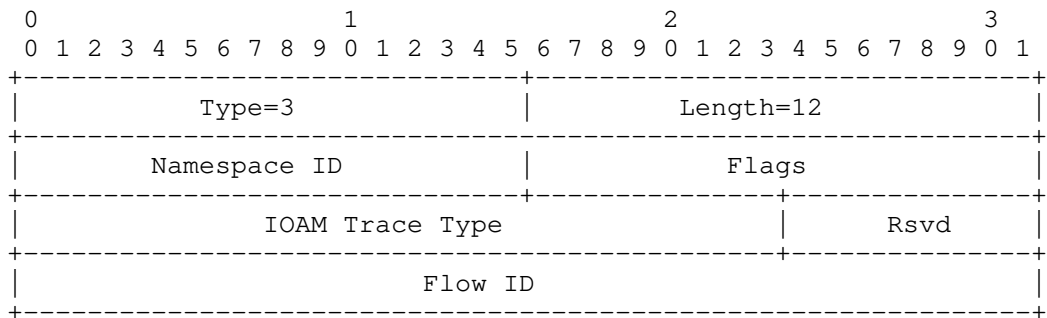


Fig. 6 IOAM Directly Export Option Sub-TLV

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:

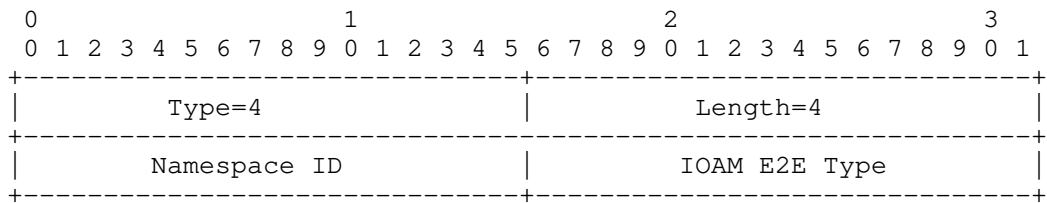


Fig. 7 IOAM Edge-to-Edge Option Sub-TLV

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-Namespaces. 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:

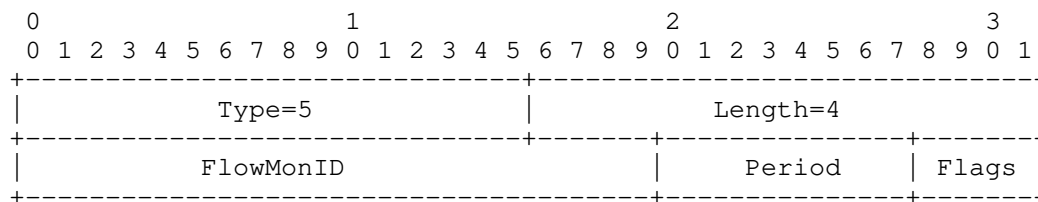


Fig. 8 Enhanced Alternate Marking Sub-TLV

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., Negl, 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: wangyalil1@huawei.com

Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: November 26, 2021

H. Bidgoli, Ed.
Nokia
V. Voyer
Bell Canada
S. Rajarathinam
Nokia
E. Hemmati
Cisco System
T. Saad
Juniper Networks
S. Sivabalan
Ciena
May 25, 2021

PCEP extensions for p2mp sr policy
draft-hsd-pce-sr-p2mp-policy-03

Abstract

SR P2MP policies are set of policies that enable architecture for P2MP service delivery. This document specifies extensions to the Path Computation Element Communication Protocol (PCEP) that allow a stateful PCE to compute and initiate P2MP paths from a Root to a set of Leaves.

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 November 26, 2021.

Copyright Notice

Copyright (c) 2021 IETF Trust and the persons identified as the document authors. All rights reserved.

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

Table of Contents

1. Introduction	3
2. Conventions used in this document	4
3. Overview of PCEP Operation in SR P2MP Network	4
3.1. High level view of P2MP Policy Objects	5
3.1.1. Shared Tree vs Non-Shared Replication Segment	6
3.2. Existing drafts used for defining a P2MP Policy	7
3.2.1. Existing Documents used by this draft	7
3.2.2. P2MP Policy Identification	8
3.2.3. Replication Segment Identification	9
3.2.4. PCECC Use in Replication Segment	9
3.3. High Level Procedures for P2MP SR LSP Instantiation	9
3.3.1. PCE-Init Procedure	9
3.3.2. PCC-Init Procedure	10
3.3.3. Common Procedure	10
3.3.4. Global Optimization of the Candidate Path	11
3.3.5. Fast Reroute	12
3.3.6. Connecting Replication Segment via Segment List	13
3.4. SR P2MP Policy and Replication Segment TLVs and Objects	13
3.4.1. SR P2MP Policy Objects	13
3.4.2. Replication Segment Objects	14
3.4.3. P2MP Policy and Replication Segment general considerations	14
4. Object Format	15
4.1. Open Message and Capability Exchange	15
4.1.1. PCECC Path Setup Capability	15
4.1.2. Association Type Capability	16
4.2. Symbolic Name in PCInit Message from PCC	16
4.3. P2MP Policy Specific Objects and TLVs	16
4.3.1. P2MP Policy Association Group for P2MP Policy	16
4.3.1.1. P2MP SR Policy Association Group Policy Identifiers TLV	16
4.3.1.2. P2MP SR Policy Association Group Candidate Path Identifiers TLV	17
4.3.1.3. P2MP SR Policy Association Group Candidate Path Attributes TLV	18
4.3.2. P2MP-END-POINTS Object	18

4.4. P2MP Policy and Replication Segment Identifier Object and TLV	21
4.4.1. Extension of the LSP Object, SR-P2MP-LSPID-TLV	21
4.5. Replication Segment	22
4.5.1. The format of the replication segment message	23
4.5.2. PCECC	23
4.5.3. Label action rules in replicating segment	26
4.5.4. SR-ERO Rules	27
4.5.4.1. SR-ERO subobject changes	27
5. Tree Deletion	28
6. Fragmentation	28
7. Example Workflows	28
8. IANA Consideration	33
9. Security Considerations	34
10. Acknowledgments	34
11. References	34
11.1. Normative References	34
11.2. Informative References	34
Authors' Addresses	35

1. Introduction

The draft [draft-ietf-pim-sr-p2mp-policy] defines a variant of the SR Policy [draft-ietf-spring-segment-routing-policy] for constructing a P2MP segment to support multicast service delivery.

A Point-to-Multipoint (P2MP) Policy connects a Root node to a set of Leaf nodes, optionally through a set of intermediate replication nodes. A Replication segment [draft-ietf-spring-sr-replication-segment], which corresponds to the state of a P2MP segment on a particular node which provide forwarding instructions for the segment.

A P2MP Policy is relevant on the root of the P2MP Tree and it contains candidate paths. The candidate paths are made of path-instances and each path-instance is constructed via replication segments. These replication segments are programmed on the root, leaves and optionally intermediate replication nodes.

A replication segments MAY be connected directly, or they MAY be connected or steered via unicast SR segment or a segment list.

For a P2MP Tree, a controller may be used to compute paths from a Root node to a set of Leaf nodes, optionally via a set of replication nodes. A packet is replicated at the root node and optionally on Replication nodes towards each Leaf node.

There are two types of a P2MP Tree: Spray and Replication.

A Point-to-Multipoint service delivery could be via Ingress Replication, known as Spray. The root unicasts individual copies of traffic to each leaf. The corresponding P2MP Policy consists of replication segments only for the root and the leaves and they are connected via a unicast SR Segment.

A Point-to-Multipoint service delivery could also be via Downstream Replication, known as Replication. The root and some downstream replication nodes replicate the traffic along the way as it traverses closer to the leaves.

The leaves and the root can be explicitly configured on the PCE or PCC can update the PCE with the information of the discovered root and leaves. As an example Multicast protocols like MVPN procedures [RFC6513] or PIM can be used to discovery the leaves and roots on the PCC and update the PCE with these relevant information. The controller can calculate the P2MP Policy and any of its associated replication segments with these info.

This document defines PCEP objects, TLVs and the procedures to instantiate a P2MP Policy and Replication Segments.

2. Conventions used in this document

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

3. Overview of PCEP Operation in SR P2MP Network

After discovering the root and the leaves on the PCE (via different mechanism mentioned in previous sections), the PCE computes the P2MP Tree and identifying the relevant Replication routers, then it programs the PCCs with relevant information needed to create a P2MP Tree.

As per draft [draft-ietf-pim-sr-p2mp-policy] a P2MP Policy is defined by Root-ID, Tree-ID and a set of leaves. A P2MP policy is a variant of SR policy as such it uses the same concept as draft [draft-ietf-pce-segment-routing-policy-cp]. A P2MP policy is composed of a collection of SR P2mp Candidate Paths. Candidate paths are computed by the PCE and can be used for P2MP Tree redundancy. Only a single candidate path may be active at each time. Each candidate paths can be globally optimized, therefore it is consists of multiple path-instances. A path-instance can be considered to a P2MP LSP. If a candidate path needs to be globally optimized two path-instances can be programmed on the root node and via make before break procedures the candidate path can be switched from path-

instance 1 to the 2nd path-instance. The forwarding states of these path-instances are build via replication segments, in short each path-instance initiated on the root has its own set of replication segments on the Root, Transit and Leaf nodes.

A replication segment is set of forwarding instructions on a specific node. Each instruction may be a PUSH or SWAP operation before forwarding out of an interface, or a POP action on bud and leaf nodes.

PCE could also calculate and download additional information for the replication segments, such as protections next-hops for link protection (FRR).

3.1. High level view of P2MP Policy Objects

- o SR P2MP Policy

- * Is only relevant on the Root of the P2MP path and is a policy on PCE. It is downloaded only on the rootnode and is identified via <Root-ID, Tree-ID> It contains the following information:

- + Root node of the P2MP Segment
 - + Leaf nodes of the P2MP Segment
 - + Tree-ID, which is a unique identifier of the P2MP tree on the Root
 - + A set of Candidate paths belonging to the policy
 - + Optional Constraints used to build these candidate paths

- o Candidate Path:

- * Is used for P2MP Tree redundancy where the candidate path with the highest preference is the active path.
 - * It can contain two path-instance for global optimization procedures (i.e. make before break)
 - * Contains information regarding protocol-id, originator, discriminator, preference, path-instances

- o Replication Segment:

- * Is the forwarding information needed on each node for building the forwarding path for each path-instance of the P2MP Candidate path.
- * Explained further in upcoming sections, there are 2 ways to identify the replication segment, depending if they are shared and non-shared
 - + It is identified via Tree-ID and Root-ID and path-instance for non-shared replication segment.
 - + It is identified via Node-ID, Replication-ID, for shared replication segment
 - + Contains forwarding instructions, in the form of a list of outgoing segments each of which may be a list
 - + On the forwarding plane the Replication Segment is identified via the incoming Replication SID.
 - + Replication segment information is downloaded on Root, Transit and Leaf nodes respectively.

3.1.1. Shared Tree vs Non-Shared Replication Segment

A non-shared Replication Segment is used when the label field of the PMSI Tunnel Attribute (PTA) is set to zero as per [draft-parekh-bess-mvpn-sr-p2mp]. This is used when there is no upstream assigned label in the PTA (provider tunnel attribute) and aggregate of MVPNs into a single P-Tunnel is not desired.

An alternative shared Replication Segment is used when the label field of the PTA is not set to Zero and there is an upstream assigned label in the PTA. In this case multiple MVPNs (VRFs) can be aggregate into a single Provider Tunnel and the upstream assigned label distinguishes the MVPNs context.

It should be noted that the shared Replication Segment can also be used to build a bypass tunnel for the purpose of fast re-route. This might be desirable if the bypass tunnel is build via the PCE and downloaded to the PCC for link protection. In doing so, multiple non-shared Replication Segments can use the shared replication segment as their bypass tunnel for link protection. The replication segments used in this bypass tunnel should only create a unicast bypass tunnel to protect the link between two replication segments on the primary path.

3.2. Existing drafts used for defining a P2MP Policy

This document attempts to leverage existing IETF draft and RFC documents which define PCEP objects, to update the PCE with Root and Leaves information when PCC Initiated method is used. Similarly, existing documents are utilized where feasible to update the PCC with relevant information to build the P2MP Policy and its Replication Segments. This document introduces new TLVs and Objects specific to a programming P2MP policy and its replication segment.

3.2.1. Existing Documents used by this draft

- o [RFC8231] The bases for a stateful PCE, and reuses the following objects or a variant of them
 - * <SRP Object>
 - * <LSP Object>
 - * A variation of the LSP identifier TLV is defined in this draft, to support P2MP LSP Identifier
- o [RFC8236] P2MP capabilities advertisement
- o [draft-ietf-pce-segment-routing-policy-cp] Candidate paths for P2MP Policy is used for Tree Redundancy. As an example, a P2MP Policy can have multiple candidate paths. Each protecting the primary candidate path. The active path is chosen via the preference of the candidate path.
- o [RFC3209] Defines the instance-ID, instance-ID is used for global optimization of a candidate path with in a P2MP policy. Each Candidate path can have 2 path-instances. These path-instances are equivalent to sub-lsps (instance-IDs). There are used for MBB and global optimization procedures. instance-ID is equivalent to LSP ID
- o [draft-ietf-spring-segment-routing-policy] Segment-list, used for connecting two non-adjacent replication policy via a unicast binding SID or Segment-list.
- o [RFC8306] P2MP End Point objects, used for the PCC to update the PCE with discovered Leaves.
- o [draft-ietf-pce-pcep-extension-for-pce-controller] for programming and identifying the Replication Segment. A new PCE CC Capability sub Tlv is introduced to indicated the support to handle PCE CC based label download for SR P2MP.

- o [draft-ietf-pce-multipath] Forwarding instruction for a P2MP LSP is defined by a set of SR-ERO sub-objects in the ERO object, ERO-ATTRIBUTES object and MULTIPATH-BACKUP TLV as defined in this draft.
- o [RFC8664] SR-ERO Sub Object used in the multipath.

It should be noted that the [draft-dhs-spring-sr-p2mp-policy-yang] can provide further details of the high level P2MP Policy Model.

3.2.2. P2MP Policy Identification

A P2MP Policy and its candidate path can be identified on the root via the P2MP LSP Object. This Object is a variation of the LSP ID Object defined in [RFC8231] and is as follow:

- o PLSP-ID: [RFC8231], is assigned by PCC and is unique per candidate path. It is constant for the lifetime of a PCEP session. Stand-by candidate paths will be assigned a new PLSP-ID by PCC. Stand-by candidate paths can co-exist with the active candidate path.
 - * Note: Every candidate path in the SR-P2MP Policy is unique with its own unique PLSP-ID and Instance-ID. But the same Tree-ID is used for all candidate paths as they are part of the same P2MP Tree.
- o Root-ID: is equivalent to the first node on the P2MP path, as per [RFC3209], Section 4.6.2.1
- o Tree-ID: is equivalent to Tunnel Identifier color which identifies a unique P2MP Policy at a ROOT and is advertised via the PTA in the BGP AD route or can be assigned manually on the root. Tree-ID needs to be unique on the root.
- o Instance-ID: LSP ID Identifier as defined in RFC 3209, is the path-instance identifier and is assigned by the PCC. As it was mentioned the candidate path can have up to two path-instance for global optimization. Note that the Root-ID, Tree-ID and Instance-ID are part of a new SR- P2MP-LSP-IDENTIFIER TLV which will be identified in this draft.
 - * Note: each Path-instance on the Root node is assigned a unique Instance-ID

3.2.3. Replication Segment Identification

The key to identify a replication segment is also a P2MP LSP Object. With varying encoding rules for the SR-P2MP-LSP- IDENTIFIER TLV which will be explained in later sections.

3.2.4. PCECC Use in Replication Segment

PCECC and a variant of CCI object is used in Replication Segment to identify a cross connect. A cross connect is a incoming SID and set of outgoing interfaces and their corresponding SID. The CCI objects contains the incoming SID while the outgoing interfaces are presented via the ERO objects, which each may contain a list of segments.

3.3. High Level Procedures for P2MP SR LSP Instantiation

A P2MP policy can be instantiated via the PCC or the PCE depending on how the root and the leaves are discovered. This document describes two way to discover the root and the leaves:

- o They can be configured and identified on the controller and are considered PCE initiated.
- o They can be discovered on the PCC via MVPN procedures [RFC6513] or legacy multicast protocols like PIM or IGMP etc... and are considered PCC initiated.

3.3.1. PCE-Init Procedure

- o PCE is informed of the P2MP request through its API or configuration mechanism to instantiate a P2MP tunnel.
- o PCE will initiate the P2MP Policy for the request, by sending a PCInitiate message to the Root.
- o Root in response to the PCInitiate message, will generate PLSP-ID for the candidate paths and an Instance-ID for the Path-Instance (LSP-ID) contained with in the candidate path. The tree-id for the P2MP Policy is also filled. PCC will reports back the PLSP-ID, Instance-ID and tree-id via PCRpt message
 - * Optionally, the Root can add any additional leaves that were discovered by multicast procedures in this PCRpt message.
- o PCE will send a PCInitiate message to the Root, Transit and the Leaf nodes to download the Replication Segment information. These messages will utilize the CCI object to encode the forwarding instruction information.

- o PCE will then send a PCUpdate to the root indicating the association information (Candidate path) , and implicitly indicate it to bind to the latest CCI information downloaded.

3.3.2. PCC-Init Procedure

After Root (PCC) discovers the leaves (as an example via MVPN Procedures or other mechanism), the following communication happens between the PCE and PCCs

- o Root sends a PCRpt message for P2MP policy to PCE including the Root-ID, Tree-ID, PLSP-ID, Instance-ID, symbolic-path-name, and any leaves discovered until then.
- o PCE on receiving of this report, will compute the P2MP Policy and its replication segments.
 - * PCE will send a PCInitiate message to the Root, Transit and the Leaf nodes to download the Replication Segment information. These messages will utilize the CCI object to encode the forwarding instruction information.
 - * PCE will then send a PCUpdate to the root indicating the association information (Candidate path) , and implicitly indicate it to bind to the latest CCI information downloaded.

3.3.3. Common Procedure

The following procedures are the same for PCE or PCC Init.

- o PCE will download the replication segments for the Candidate-path's path-instances to all the leaves and transit nodes using PCInitiate message with PLSP-ID = 0, Instance-ID =0, symbolic path name, Root-address, Tree-id(assigned by the root). This PCInitiate message includes the EROs needed for the replication segments. These messages will utilize the CCI object to encode the forwarding instruction information.
- o Any new candidate path for the P2MP Policy is downloaded by PCE to the Root by sending a PCInitiate message
 - * it should be noted, PLSP-ID, Path-Instance ID and the Tree-ID are generated by the PCC for these new candidate paths and their Path-instances
 - * Any update to the Candidate Paths or Replication Segments is done via the PCUpd message. Association object need to be

present for Candidate Path updates and CCI object for the replication segment updates.

- o The PCE will also download the necessary replication segment for the candidate path and its path-instances to the root, leaves and the transit nodes via a PCInit message
- o New leaves can be discovered via Multicast procedures, and new replication segments can be instantiated or existing one updated to reach these leaves
 - * If these leaves reside on routers that are part of the P2MP LSP path, then PCUpd is sent from PCE to necessary PCCs (LEAVES, TRANSIT or ROOT) with the correct PLSP-ID, Instance-ID, Tree-ID and CC-ID.
 - * If the new leaves are residing on routers that are not part of the P2MP Tree yet, then a PCInitiate message is sent down with PLSP-ID=0 and Instance-ID=0 on the corresponding routers.
- o The active candidate-path is indicated by the PCC through the operational bits(Up/Active) of the LSP object in the PCRpt message. If a candidate path needs to be removed, PCE sends PC Initiate message, setting the R-flag in the LSP object and R bit in the SRP-object.
- o To remove the entire P2MP-LSP, PCE needs to send PCInitiate remove messages for every candidate path of the P2MP POLICY to the root and send PCInitiate remove messages for every Replication Regment on all the PCCs on the P2MP Tree. The R bit in the LSP Object as defined in [RFC8231], refers to the removal of the LSP as identified by the SR-P2MP-POLICY-ID-TLV (defined in this document). An all zero (SR-P2MP-LSP-ID-TLV defines to remove all the state of the corresponding PLSP-ID.
- o A candidate path is made active based on the preference of the path. If the Root is programed with multiple candidate paths from different sources, as an example PCE and CLI, based on its tie-breaking rules, if it selects the CLI path, it will send a report to PCE for the PCE path indicating the status of label-download and sets operational bit of the LSP object to UP and Not Active . At any instance, only one path will be active

3.3.4. Global Optimization of the Candidate Path

When a P2MP LSP needs to be optimized for any reason (i.e. it is taking a FRR tunnel or new routers are added to the network) a global optimization of the candidate path is possible.

Each Candidate Path can contain two Path-Instances. The current unoptimized Path-Instance is the active instance and its replication segments are forwarding the multicast PDUs from the root to the leaves. However the second optimized Path-Instance will be setup with its own unique replication segments throughout the network, from the Root to the leaves. These two Path-Instances can co-exist. The two Path-Instances are uniquely identified by their Instance-ID in the SR-P2MP-POLICY-ID-TLV (defined in this document). After the optimized LSP has been downloaded successfully PCC MUST send two reports, reporting UP of the new path indicating the new LSP-ID, and a second reporting the tear down of the old path with the R bit of the LSP Object SET with the old Instance-ID in the SR-P2MP-POLICY-ID-TLV. This MBB procedure will move the multicast PDUs to the optimized Path-Instance.

The leaf should be able to accept traffic from both Path-Instances to minimize the traffic outage by the Make Before Break process.

3.3.5. Fast Reroute

Currently this draft identifies the Facility FRR procedures. In addition, only LINK Protection procedures are defined. How the Facility Path is built and instantiated is beyond the scope of this document.

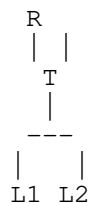


Figure 1

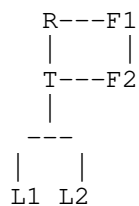


Figure 2

As an example, the bypass path (unicast bypass) between the PLR and MP can be constructed via SR or even via a shared tree (replication segment).

As an example, in figure 1 the detour path between R and T is the 2nd fiber between these nodes. As such the bypass path could be setup on the 2nd fiber. That said in figure 2 the bypass path is traversing multiple nodes and this example a unicast SR path might be ideal for setting up the detour path.

In addition, PHP procedure and implicit null label on the bypass path can be implemented to reduce the PCE programming on the MP PCC.

Optional shared replication segments can be used in networks that do not have unicast SR turned on. These shared replication segments can be programmed on the bypass nodes without a P2MP Policy. The replication segments on primary path can use these shared replication segments as a protection tunnel to protect links.

3.3.6. Connecting Replication Segment via Segment List

There could be nodes between two replication segment that do not support P2MP Policy or Replication segment. It is possible to connect two non-adjacent Replication segments via a unicast segment routing path via a SID list, comprised of any IGP supported segment types (ex: Binding, Adjacency, Node) to forward to the next replicating node. This information is encoded via the SR-ERO sub-objects and ERO-attributes objects. The last segment in an encoding SID list MUST be a replication segment

3.4. SR P2MP Policy and Replication Segment TLVs and Objects

3.4.1. SR P2MP Policy Objects

SR P2MP Policy can be constructed via the following objects

<Common Header>

<SRP>

<P2MP LSP>

[<association-list>]

optionally if the root is updating the PCE with end point list the end-point-list object can be added.

[<end-points-list>]

3.4.2. Replication Segment Objects

Replication segment can be constructed via the following objects

```
<Common Header>
<SRP>
<P2MP LSP>
(<cci-list>|
<CCI><intended-path>))
<cci-list> ::= <CCI>
               [<cci-list>]
<intended-path> ::= ((<PATH-ATTRIB><ERO>)
                    [<intended-path>])
```

Path-attribute as per [draft-ietf-pce-multipath]

3.4.3. P2MP Policy and Replication Segment general considerations

The above new objects and TLV's defined in this document can be included in PCRpt, PCInitiate and PCUpd messages.

It should be noted that every PCRpt, PCInitiate and PCUpd messages will contain full list of the Leaves and segment and forwarding information that is needed to build the Candidate path and its Replication segments. They will never send the delta information related to the new leaves or forwarding information that need to be added or updated. This is necessary to ensure that PCE or any new PCE is in sync with the PCC.

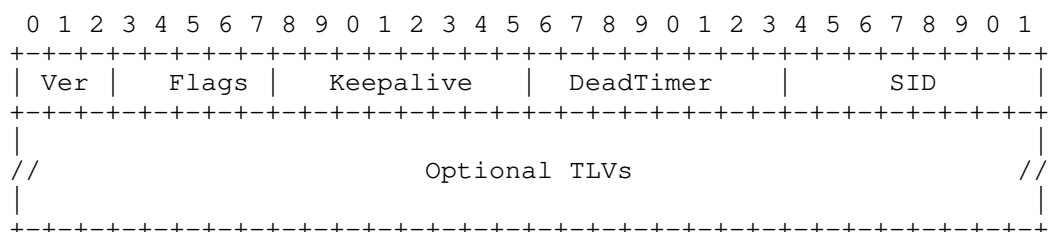
When a PCRpt, PCInitiate and PCUpd messages is sent via PCEP it maintains the previous ERO Path IDs and generates new Path IDs for new instructions, as per [draft-ietf-pce-multipath]. The PATH IDs are maintained for each specific forwarding instructions until the instructions are deleted. For example: When the first leaf is added, the PCE will update with PathID 1 to the PCC. When the second leaf is added, according to the path calculated, PCE might just append the existing instruction Path ID 1 with a new Path ID 2 to construct the new PCUpd message.

The CCI Object is used to identify the entire cross connect of incoming segment and the set of outgoing Interfaces and their corresponding SIDs/SIDList. Any modification to the cross connect should use this CCI ID to identify the cross connect uniquely. Leaves and their corresponding Path IDs can be removed from the cross connect identified via the CCI. The CC-ID is assigned by the PCE.

4. Object Format

4.1. Open Message and Capability Exchange

Format of the open Object:



All the nodes need to establish a PCEP connection with the PCE.

During PCEP Initialization Phase, PCEP Speakers need to set flags N, M, P in the STATEFUL-PCE-CAPABILITY TLV as defined in [draft-ietf-pce-stateful-pce-p2mp] section-5.2

This draft extends the PCEP OPEN object by defining an optional TLV to indicate the PCE's capability to perform SR-P2MP path computations with a new IANA capability type.

The inclusion of this TLV in an OPEN object indicates that the sender can perform SR-P2MP path computations. This will be similar to the P2MP-CAPABILITY defined in [RFC8306] section-3.1.2 and a new value needs to be defined for SR-P2MP.

4.1.1. PCECC Path Setup Capability

A PST of PCECC is also added as per [draft-ietf-pce-pcep-extension-for-pce-controller].

This document also introduces a new bit S in the SR PCECC capacity Sub TLV indicating the support to handle PCECC based label download for Replication segment.



Also, the N,M,P bits in STATEFUL-PCE-CAPABILITY TLV should be SET.

4.1.2. Association Type Capability

A Assoc-Type-List TLV as per [RFC8697] section 3.4 should be send via PCEP open object with following association type

Association Type Value	Association Name	Reference
TBD1	P2MP SR Policy Association	This document

OP-CONF-Assoc-RANGE (Operator-configured Association Range) should not be set for this association type and must be ignored.

The open message MUST include the MULTIPATH CAPABILITY TLV as defined in [draft-ietf-pce-multipath]

4.2. Symbolic Name in PCInit Message from PCC

As per [RFC8231] section 7.3.2. a Symbolic Path Name TLV should uniquely identify the P2MP path on the PCC. This symbolic path name is a human-readable string that identifies an P2MP LSP in the network. It needs to be constant through the lifetime of the P2MP path.

As an example in the case of P2MP LSP the symbolic name can be p2mp policy name + candidate path name of the LSP.

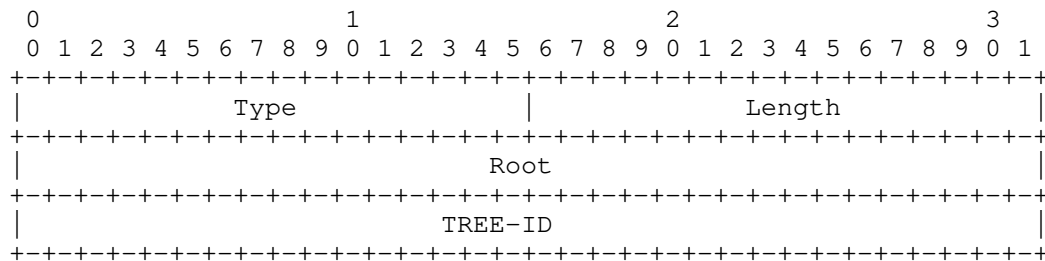
4.3. P2MP Policy Specific Objects and TLVs

4.3.1. P2MP Policy Association Group for P2MP Policy

Two ASSOCIATION object types for IPv4 and IPv6 are defined in [RFC8697]. The ASSOCIATION object includes "Association type" indicating the type of the association group. This document adds a new Association type. Association type = TBD1 "P2MP SR Policy Association Type" for SR Policy Association Group (P2MP SRPAG). As per [draft-barth-pce-segment-routing-policy-cp] section 5, three new TLVs are identified to carry association information: P2MP-SRPAG-POL-ID-TLV, P2MP-SRPAG-CPATH-ID-TLV, P2MP-SRPAG-CPATH-ATTR-TLV

4.3.1.1. P2MP SR Policy Association Group Policy Identifiers TLV

The P2MP-SRPOLICY-POL-ID TLV is a mandatory TLV for the P2MP-SRPAG Association. Only one P2MP-SRPOLICY-POL-ID TLV can be carried and only the first occurrence is processed and any others MUST be ignored.



Type: TBD2 for "P2MP-SR-POLICY-POL-ID" TLV.

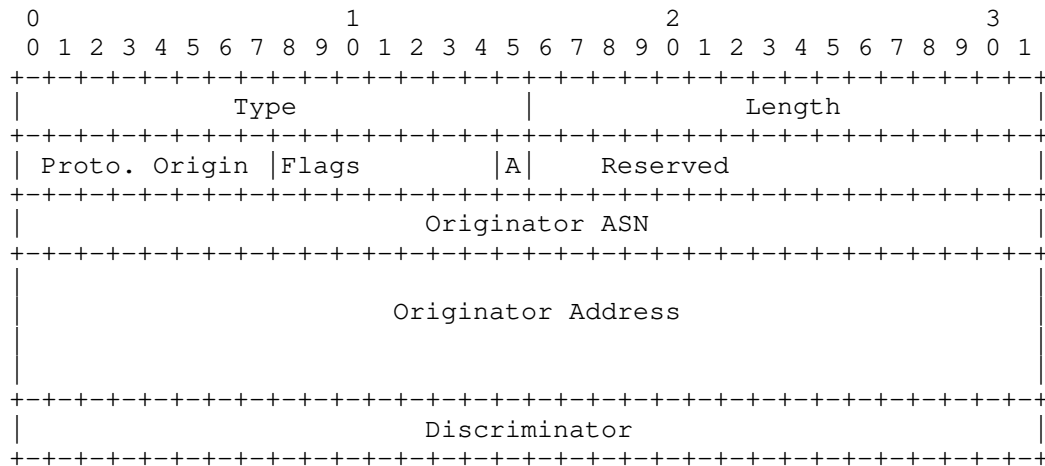
Length: 8 or 20, depending on length of End-point (IPv4 or IPv6)

Tunnel Sender Address : Can be either IPv4 or IPv6, this value is the value of the root loopback IP.

Tree-ID: Tree ID that the replication segment is part of as per draft-ietf-spring-sr-p2mp-policy

4.3.1.2. P2MP SR Policy Association Group Candidate Path Identifiers TLV

The P2MP-SRPOLICY-CPATH-ID TLV is a mandatory TLV for the P2MPSRPAG Association. Only one P2MP-SRPOLICY-CPATH-ID TLV can be carried and only the first occurrence is processed and any others MUST be ignored.



Type: TBD3 for "P2MP-SR-POLICY-CPATH-ID" TLV.

Length: 28.

Protocol Origin: 8-bit value that encodes the protocol origin, as specified in [I-D.ietf-spring-segment-routing-policy] Section 2.3.

Flags : A: This candidate path is active. At any instance only one candidate path can be active. PCC indicates the active candidate path to PCE through this bit. Reserved: MUST be set to zero on transmission and ignored on receipt.

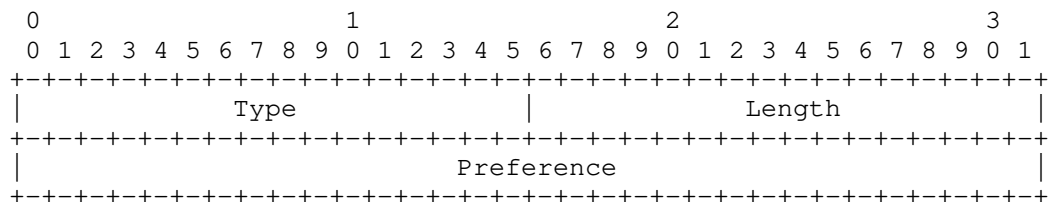
Originator ASN: Represented as 4 byte number, part of the originator identifier, as specified in [draft-ietf-spring-segment-routing-policy] Section 2.4.

Originator Address: Represented as 128 bit value where IPv4 address are encoded in lowest 32 bits, part of the originator identifier, as specified in [draft-ietf-spring-segment-routing-policy] Section 2.4.

Discriminator: 32-bit value that encodes the Discriminator of the candidate path.

4.3.1.3. P2MP SR Policy Association Group Candidate Path Attributes TLV

The P2MP-SRPOLICY-CPATH-ATTR TLV is an optional TLV for the SRPAG Association. Only one P2MP-SRPOLICY-CPATH-ATTR TLV can be carried and only the first occurrence is processed and any others MUST be ignored.



Type: TBD4 for "P2MP-SRPOLICY-CPATH-ATTR" TLV.

Length: 4. Preference: Numerical preference of the candidate path, as specified in [draft-ietf-spring-segment-routing-policy] Section 2.7.

If the TLV is missing, a default preference of 100 as specified in [draft-ietf-spring-segment-routing-policy] is used.

4.3.2. P2MP-END-POINTS Object

In order for the Root to indicate operations of its leaves (Add/Remove/Modify/DoNotModify), the PC Report message is

extended to include P2MP End Point <P2MP End-points> Object which is defined in [RFC8306]

The format of the PC Report message is as follow:

<Common Header>

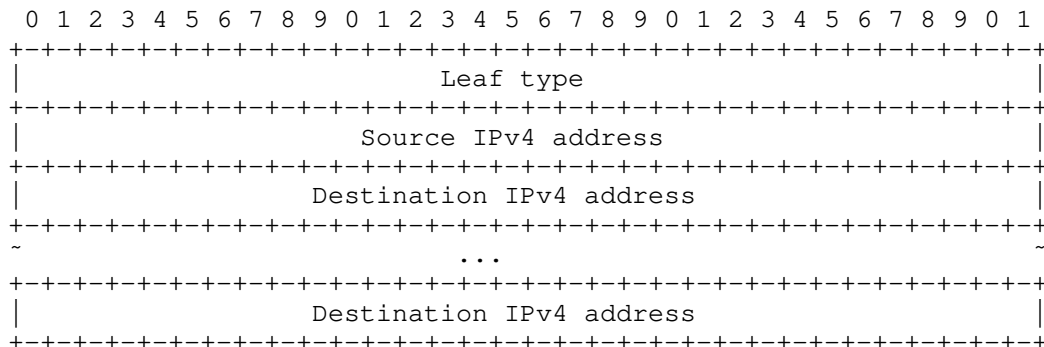
[<SRP>]

<LSP>

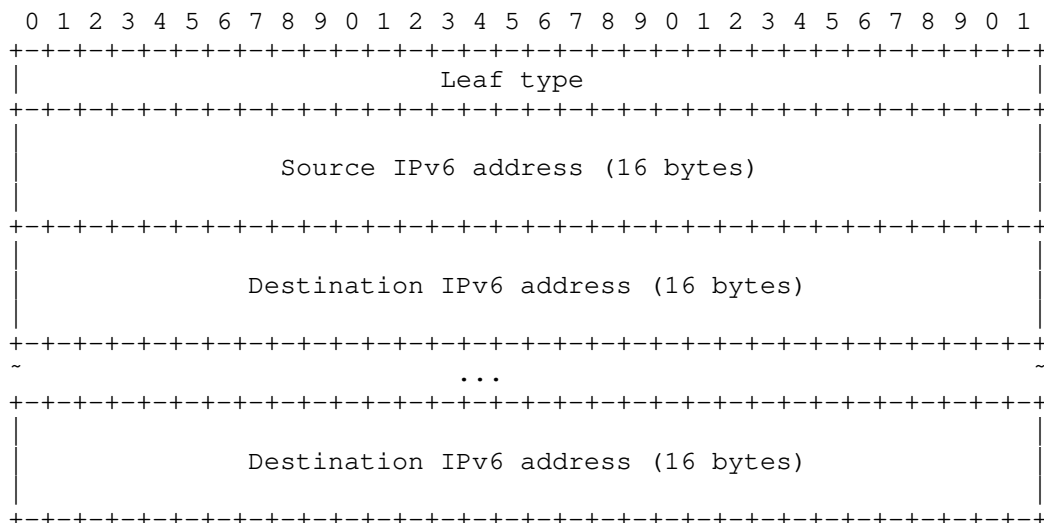
[<association-list>]

[<end-points-list>]

IPv4-P2MP END-POINTS:



IPv6-P2MP END-POINTS:



Leaf Types (derived from [RFC8306] section 3.3.2) :

1. New leaves to add (leaf type = 1)
2. Old leaves to remove (leaf type = 2)
3. Old leaves whose path can be modified/reoptimized (leaf type = 3), Future reserved not used for tree SID as of now.
4. Old leaves whose path must be left unchanged (leaf type = 4)

5. the entire pce leaf list is overwritten and replaced with the new leaf list (leaf type = 5)

A given P2MP END-POINTS object gathers the leaves of a given type. Note that a P2MP report can mix the different types of leaves by including several P2MP END-POINTS objects. The END-POINTS object body has a variable length. These are multiples of 4 bytes for IPv4, multiples of 16 bytes, plus 4 bytes, for IPv6.

4.4. P2MP Policy and Replication Segment Identifier Object and TLV

As it was mentioned previously both P2MP Policy and Replication Segment are identified via the LSP object and more precisely via the SR-P2MP-LSPID-TLV

The P2MP Policy uses the PLSP-ID to identify the Candidate Paths and the Instance-ID to identify a Path-Instance within the Candidate path.

On the other hand the Replication Segment uses the SR-P2MP-LSPID-TLV to identify and correlate a Replication Segment to a P2MP Policy

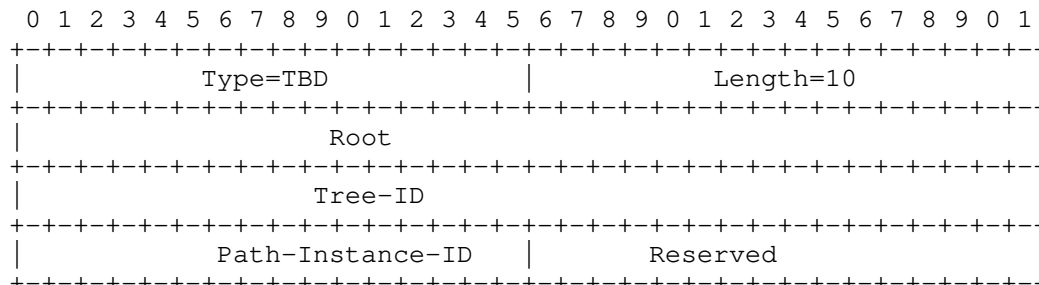
As it was noted previously on the Root, the P2MP Policy and the Replication Segment is downloaded via the same PCUpd message.

4.4.1. Extension of the LSP Object, SR-P2MP-LSPID-TLV

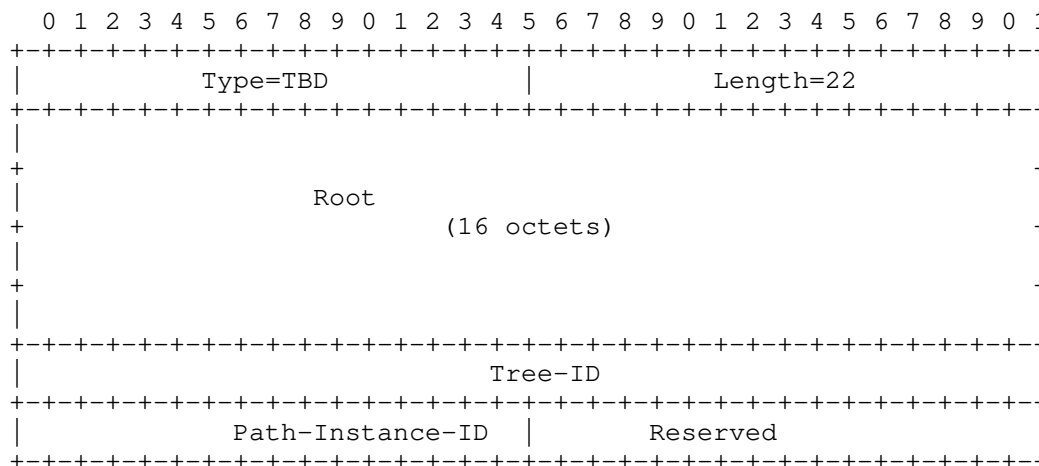
The LSP Object is defined in Section 7.3 of [RFC8231]. It specifies the PLSP-ID to uniquely identify an LSP that is constant for the life time of a PCEP session. Similarly, for a P2MP tunnel, the PLSP-ID identifies a Candidate Path uniquely within the P2MP policy.

The LSP Object MUST include the new SR-P2MP-POLICY-ID-TLV (IPv4/IPv6) defined in this document below. This is a variation to the P2MP object defined in [draft-ietf-pce-stateful-pce-p2mp]

SR-IPV4-P2MP-POLICY-ID TLV:



SR-IPV6-P2MP-POLICY-ID TLV :



The type (16-bit) of the TLV is TBD (need allocation by IANA).

Root: Source Router IP Address

Tree-ID: Unique Identifier of this P2MP LSP on the Root.

Instance-ID : Contains 16 Bit instance ID.

4.5. Replication Segment

As per [draft-ietf-spring-sr-replication-segment] a replication segment has a next-hop-group which MAY contain a single outgoing replication SID or a list of SIDs (sr-policy-sid-list) In either case there needs to be a replication SID at the bottom of the stack. This

means two replication segments can be directly connected or connected via a SR domain.

4.5.1. The format of the replication segment message

The format of a Replication Segment message encoding is similar to P2MP Policy. However, the P2MP Policy contains the association object and the replication segment message does not contain the association object. In addition the replication segment uses the CCI object to identify a P2MP cross connect. The replication segment is downloaded individually to the root, transit and leaf nodes without the P2MP Policy. The P2MP Policy is a Root Concept. The replication segment uses SR-P2MP-LSPID-TLV as its identifier. The TLV is coded differently for shared and non-shared case.

- o In the case of a replication segment being shared, the Tree-ID in the SR-P2MP-POLICY Identifier TLV is the replication-id of the Replication Segment and Root = 0, Instance-Id = 0. When downloading a shared replication segment from PCE through a PCEInitiate message, the SR-P2MP-POLICY Identifier TLV is all 0, and on the report back from PCC, PCC generates PLSP-ID, Replication-id (Tree-id field will be populated with replication-id). Instance-id will be 0.

4.5.2. PCECC

The CCI Object as defined in [draft-ietf-pce-pcep-extension-for-pce-controller] is used to identify a forwarding instruction in the Replication Segment. A forwarding instruction is incoming SID and a set of outgoing branches. The CCI Object-Type of 1 is used for the MPLS Label. The label in the CCI Object is the incoming SID. The outgoing SIDs are defined by the ERO Objects.

The CCI Object can be include in Reports, initiate and Update messages for Replication Segments.

The PCEInitiate message defined in [RFC8281] and extended in [draft-ietf-pce-pcep-extension-for-pce-controller] is further extended to support SR-P2MP replication segment based central control instructions.

The format of the extended PCInitiate message is as follows:

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

Where:

<Common Header> is defined in [RFC5440]

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

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

```
<PCE-initiated-lsp-central-control> ::= <SRP>
                                         <LSP>
                                         (<cci-list> |
                                         (<CCI><intended-path>))
```

```
<cci-list> ::= <CCI>
               [<cci-list>]
```

```
<intended-path> ::= ((<PATH-ATTRIB><ERO>)
                     [<intended-path>])
```

Where:

<PCE-initiated-lsp-instantiation> and
<PCE-initiated-lsp-deletion> are as per
[RFC8281].

The LSP and SRP object is defined in [RFC8231]. The <intended-path> is as per [RFC8281] [draft-ietf-pce-multipath] (PATH-ATTRIB and ERO).

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> |
                               (<CCI><intended-path>))
```

```
<cci-list> ::= <CCI>
               [<cci-list>]
```

Where:

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

The <intended-path> is as per [draft-ietf-pce-multipath] (PATH-ATTRIB and ERO).

This document extends the use of PCUpd message with SR-P2MP CCI as follows:

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

Where:

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

```
<update-request> ::= (<lsp-update-request> |
                        <central-control-update>)
```

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

```
<central-control-update> ::= <SRP>
                        <LSP>
                        (<CCI><intended-path>)
```

Where:

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

The <intended-path> is as per [draft-ietf-pce-multipath] (PATH-ATTRIB and ERO).

4.5.3. Label action rules in replicating segment

The node action and role of ingress, transit, leaf or bud, is indicated via a new Node Role TLV. This document introduces a new SR-P2MP-NODE-ROLE TLV (Type To be assigned by IANA) that will be present in the PATH-ATTRIB object.

```

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=4              |
+-----+-----+-----+-----+-----+-----+-----+-----+
|   Role Type   |                                     | Reserved         |
+-----+-----+-----+-----+-----+-----+-----+-----+
```

- o ingress, role type = 1
- o transit, role type = 2
- o leaf, role type = 3
- o bud, role type = 4

4.5.4. SR-ERO Rules

Forwarding information of a replication segment can be configured and steered via many different mechanisms.

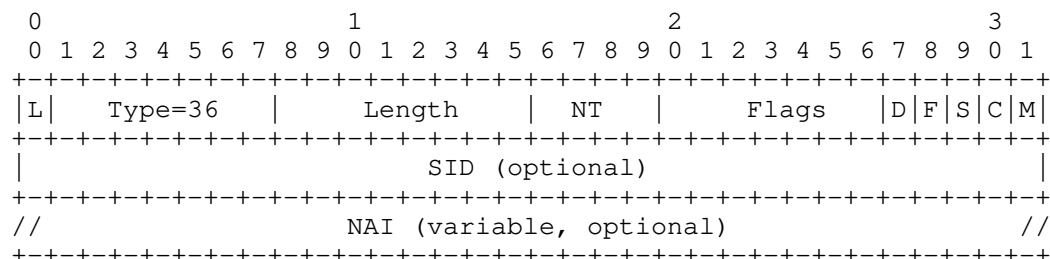
As an example a replication SID can be steered via:

1. Replication SID steered with an IPv4/IPv6 directly connected nexthop
 - * In this case there will be two SR-ERO in the ERO Object, with the Replication SID SR-ERO at the bottom and the IPv4/IPv6 SR-ERO on the top.
2. Replication SID steered with an IPv4/IPv6 loopback address that reside on the directly connected router.
 - * In this case there will be two SR-ERO in the ERO Object, with the Replication SID SR-ERO at the bottom and the IPv4/IPv6 SR-ERO on the top.
 - * In addition a new flag D is added to the SR-ERO to signal that the Loopback nexthop is connected to the directly attached router.
3. Replication SID steered with unnumbered IPv4/IPv6 directly connected Interface
4. Replication SID steered via a SR adjacency or node SID
 - * In this case even a sid-list can be used to traffic engineer the path between two Replication Segment
 - * The Replication SID SR-ERO is at the bottom while the segments describing the path are on top in order.

4.5.4.1. SR-ERO subobject changes

SR-ERO from RFC 8664 is used to construct the forwarding information needed for Replication Segment.

A new D flag was added to indicate a loopback nexthop that is residing on the directly attached router. It should be noted that this flag should be set only for the loopback case and not for a local interface as a nexthop.



Flags : F, S, C, M are already defined in rfc8664.

This document defines a new flag D: If the next-hop in NAI field is system IP or loopback, this bit indicates whether the system IP / loopback is directly connected router or not. If set indicates directly connected address. When this bit is set, F bit should be 0 (meaning NAI should be present)

5. Tree Deletion

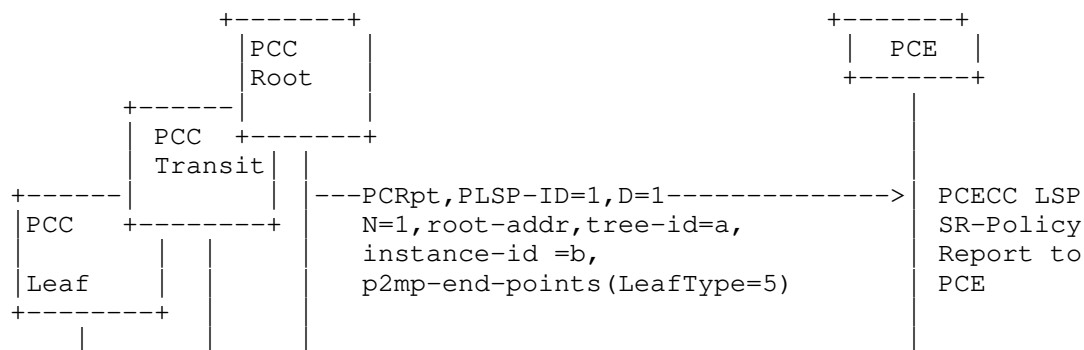
To delete the entire tree (P2MP LSP), Root send a PCRpt message with the R bit of the LSP object set and all the fields of the SR-P2MP-LSP-ID TLV set to 0(indicating to remove all state associated with this P2MP tunnel). The PCE in response sends a PCInitiate message with R bit in the SRP object SET to all nodes along the path to indicate deletion of the entries.

6. Fragmentation

The Fragmentation bit in the LSP object (F bit) can be used to indicate a fragmented PCEP message

7. Example Workflows

PCC-Initiated Workflow

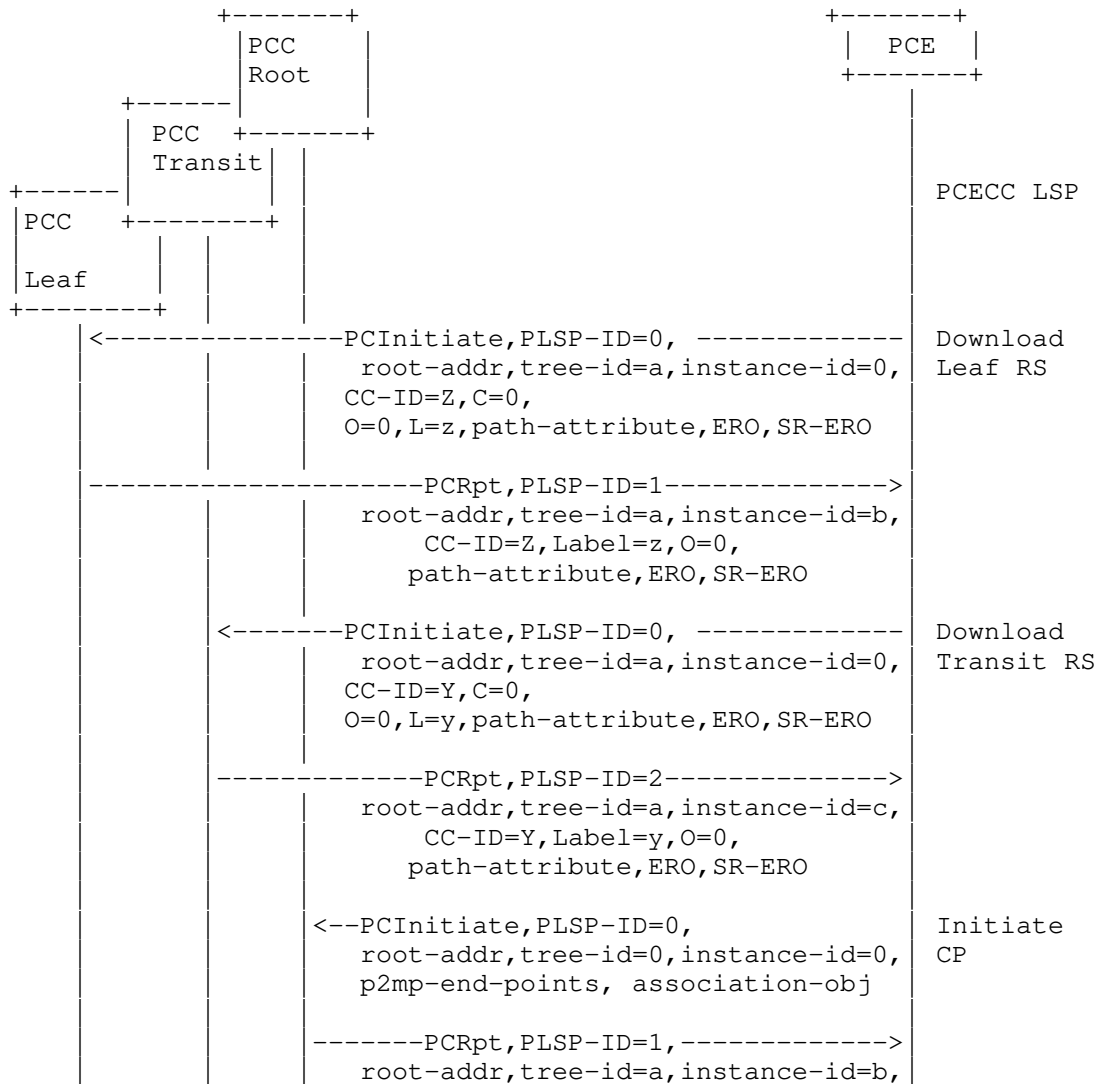


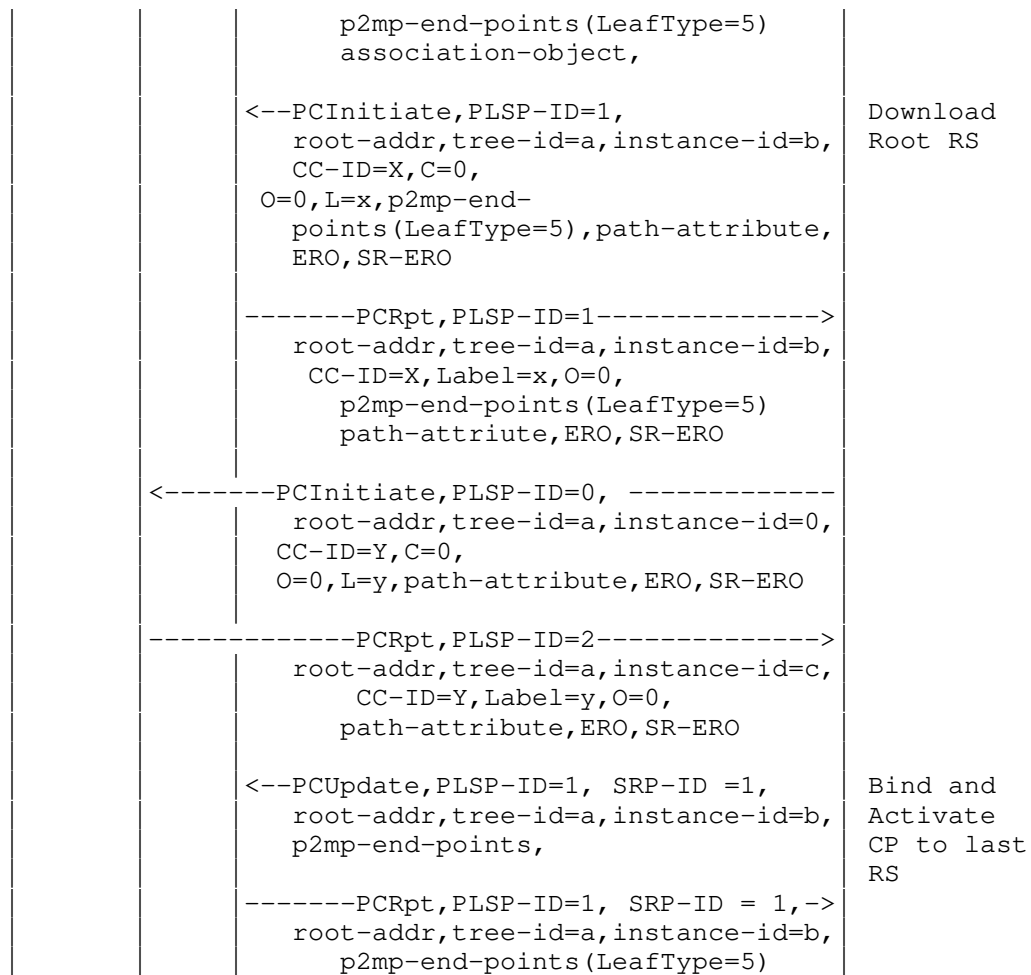
		<--PCUpdate,PLSP-ID=1, SRP-ID =1, root-addr,tree-id=a,instance-id=b, p2mp-end-points, association-obj	Update CP
		-----PCRpt,PLSP-ID=1, SRP-ID = 1,-> root-addr,tree-id=a,instance-id=b, p2mp-end-points(LeafType=5) association-object,	
<-----		PCInitiate,PLSP-ID=0, ----- root-addr,tree-id=a,instance-id=0, CC-ID=Z,C=0, O=0,L=z,path-attribute,ERO,SR-ERO	Download Leaf Replication Segment (RS)
		-----PCRpt,PLSP-ID=1-----> root-addr,tree-id=a,instance-id=b, CC-ID=Z,Label=z,O=0, path-attribute,ERO,SR-ERO	
	<-----	PCInitiate,PLSP-ID=0, ----- root-addr,tree-id=a,instance-id=0, CC-ID=Y,C=0, O=0,L=y,path-attribute,ERO,SR-ERO	Download Transit RS
		-----PCRpt,PLSP-ID=2-----> root-addr,tree-id=a,instance-id=c, CC-ID=Y,Label=y,O=0, path-attribute,ERO,SR-ERO	
		<--PCInitiate,PLSP-ID=1, root-addr,tree-id=a,instance-id=b, CC-ID=X,C=0, O=0,L=x,p2mp-end- points(LeafType=5),path-attribute, ERO,SR-ERO	Download Root RS
		-----PCRpt,PLSP-ID=1-----> root-addr,tree-id=a,instance-id=b, CC-ID=X,Label=x,O=0, p2mp-end-points(LeafType=5) path-attribute,ERO,SR-ERO	
		<--PCUpdate,PLSP-ID=1, SRP-ID =2, root-addr,tree-id=a,instance-id=b, p2mp-end-points	Activate CP to last RS
		-----PCRpt,PLSP-ID=1, SRP-ID =2, -> root-addr,tree-id=a,instance-id=b,	

p2mp-end-points (LeafType=5)

Note that on transit / leaf Initiate is with PLSP-ID = 0. Therefore PLSP-ID is locally unique to a node. It should be noted that the CC-ID does not need to be constant across all nodes that make up the path.

PCE-Initiated workflow

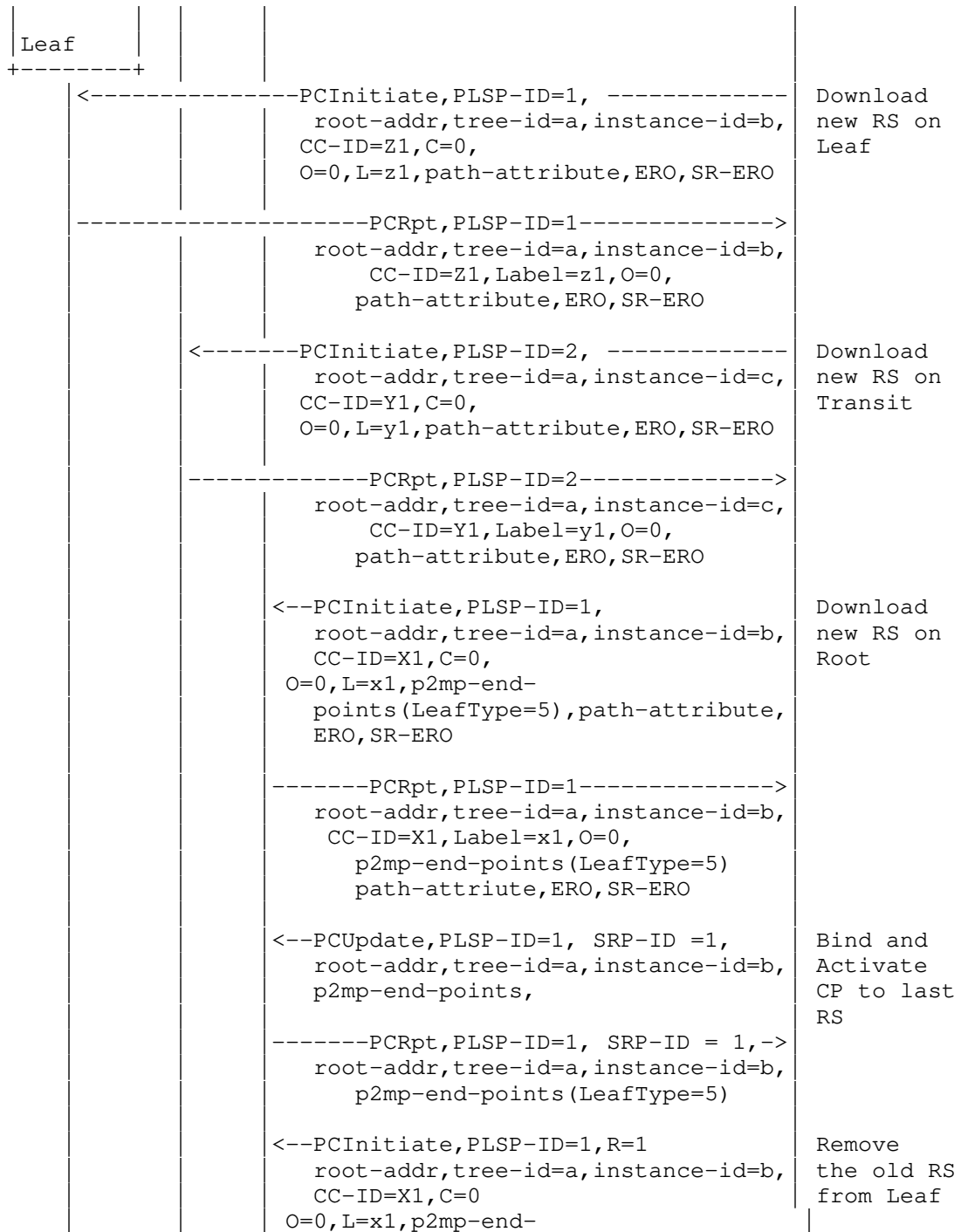




MBB Workflow:

Common (PCE-INIT, PCC-INIT) MBB





		points(LeafType=5),path-attribute, ERO,SR-ERO	
		-----PCRpt,PLSP-ID=1, R=1-----> root-addr,tree-id=a,instance-id=b, CC-ID=X1,Label=x1,O=0, p2mp-end-points(LeafType=5) path-attribute,ERO,SR-ERO	
	<-----	PCInitiate,PLSP-ID=2, R=1----- root-addr,tree-id=a,instance-id=c, CC-ID=Y1,C=0, O=0,L=y1,path-attribute,ERO,SR-ERO	Remove the old RS from Transit
		-----PCRpt,PLSP-ID=2, R=1-----> root-addr,tree-id=a,instance-id=c, CC-ID=Y1,Label=y1,O=0, path-attribute,ERO,SR-ERO	
	<-----	PCInitiate,PLSP-ID=1,R=1----- root-addr,tree-id=a,instance-id=b, CC-ID=Z1,C=0, O=0,L=z1,path-attribute,ERO,SR-ERO	Remove the old RS from Root
		-----PCRpt,PLSP-ID=1,R=1-----> root-addr,tree-id=a,instance-id=b, CC-ID=Z1,Label=z1,O=0, path-attribute,ERO,SR-ERO	

8. IANA Consideration

1. This draft extends the PCEP OPEN object by defining an optional TLV to indicate the PCE's capability to perform SR-P2MP path computations with a new IANA capability type (TBD).
2. PCEP open object with a new association type " P2MP SR Policy Association " value (TBD).
3. A new Association type. Association type = TBD1 "P2MP SR Policy Association Type" for SR Policy Association Group (P2MP SRPAG)
 1. three new TLVs are identified to carry association information: P2MP-SRPAG- POL-ID-TLV, P2MP-SRPAG-CPATH-ID-TLV, P2MP-SRPAG-CPATH-ATTR-TLV
4. Two new TLVs for Identifying the P2MP Policy and the Replication segment SR-IPV4-P2MP-POLICY-ID TLV and SR-IPV6-P2MP-POLICY-ID TLV

5. A new SR-P2MP-NODE-ROLE TLV (Type To be assigned by IANA) that will be present in the PATH-ATTRIB object

9. Security Considerations

TBD

10. Acknowledgments

The authors would like to thank Tanmoy Kundu and Stone Andrew at Nokia for their feedback and major contribution to this draft.

11. References

11.1. Normative References

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

11.2. Informative References

[draft-barth-pce-segment-routing-policy-cp]

.

[draft-dhs-spring-sr-p2mp-policy-yang]

.

[draft-ietf-pce-multipath]

.

[draft-ietf-pce-pcep-extension-for-pce-controller]

.

[draft-ietf-pce-segment-routing-policy-cp]

.

[draft-ietf-pce-stateful-pce-p2mp]

.

[draft-ietf-pim-sr-p2mp-policy]

"D. Yoyer, C. Filsfils, R.Prekh, H.bidgoli, Z. Zhang,
"draft-voyer-pim-sr-p2mp-policy"", October 2019.

[draft-ietf-spring-segment-routing-policy]

.

[draft-ietf-spring-sr-replication-segment]
"D. Yoyer, C. Filsfils, R.Prekh, H.bidgoli, Z. Zhang,
"draft-voyer-pim-sr-p2mp-policy "draft-voyer-spring-sr-
replication-segment"", July 2020.

[draft-parekh-bess-mvpn-sr-p2mp]

.

[draft-sivabalan-pce-binding-label-sid]

.

[RFC3209] .

[RFC5440] .

[RFC6513] .

[RFC8231] .

[RFC8236] .

[RFC8281] .

[RFC8306] .

[RFC8664] .

[RFC8697] .

Authors' Addresses

Hooman Bidgoli (editor)
Nokia
Ottawa
Canada

Email: hooman.bidgoli@nokia.com

Daniel Voyer
Bell Canada
Montreal
Canada

Email: daniel.yover@bell.ca

Saranya Rajarathinam
Nokia
Mountain View
US

Email: saranya.Rajarathinam@nokia.com

Ehsan Hemmati
Cisco System
San Jose
USA

Email: ehemmati@cisco.com

Tarek Saad
Juniper Networks
Ottawa
Canada

Email: tsaad@juniper.com

Siva Sivabalan
Ciena
Ottawa
Canada

Email: ssivabal@ciena.com

PCE Working Group
Internet-Draft
Intended status: Standards Track
Expires: December 9, 2021

A. Wang
China Telecom
B. Khasanov
Yandex LLC
S. Fang
R. Tan
Huawei Technologies, Co., Ltd
C. Zhu
ZTE Corporation
June 7, 2021

PCEP Extension for Native IP Network
draft-ietf-pce-pcep-extension-native-ip-14

Abstract

This document defines the Path Computation Element Communication Protocol (PCEP) extension for Central Control Dynamic Routing (CCDR) based application in Native IP network. The scenario and framework of CCDR in native IP is described in [RFC8735] and [RFC8821]. This draft describes the key information that is transferred between Path Computation Element (PCE) and Path Computation Clients (PCC) to accomplish the End to End (E2E) traffic assurance in Native IP network under central control mode.

Status of This Memo

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

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

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

This Internet-Draft will expire on December 9, 2021.

Copyright Notice

Copyright (c) 2021 IETF Trust and the persons identified as the document authors. All rights reserved.

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

Table of Contents

1. Introduction	3
2. Conventions used in this document	3
3. Terminology	3
4. Capability Advertisemnt	4
4.1. Open message	4
5. PCEP messages	4
5.1. The PCInitiate message	5
5.2. The PCRpt message	6
6. PCECC Native IP TE Procedures	7
6.1. BGP Session Establishment Procedures	7
6.2. Explicit Route Establish Procedures	9
6.3. BGP Prefix Advertisement Procedures	12
7. New PCEP Objects	13
7.1. CCI Object	13
7.2. BGP Peer Info Object	14
7.3. Explicit Peer Route Object	17
7.4. Peer Prefix Advertisement Object	19
8. End to End Path Protection	21
9. Re-Delegation and Clean up	21
10. BGP Considerations	21
11. New Error-Types and Error-Values Defined	22
12. Deployment Considerations	22
13. Security Considerations	23
14. IANA Considerations	23
14.1. Path Setup Type Registry	23
14.2. PCECC-CAPABILITY sub-TLV's Flag field	24
14.3. PCEP Object Types	24
14.4. PCEP-Error Object	24
15. Contributor	25
16. Acknowledgement	25
17. Normative References	25
Authors' Addresses	27

1. Introduction

Generally, Multiprotocol Label Switching Traffic Engineering (MPLS-TE) requires the corresponding network devices support Multiprotocol Label Switching (MPLS) or Resource ReSerVation Protocol (RSVP)/Label Distribution Protocol (LDP) technologies to assure the End-to-End (E2E) traffic performance. In Segment Routing either IGP extensions or BGP are used to steer a packet through an SR Policy instantiated as an ordered list of instructions called "segments". But in native IP network, there will be no such signaling protocol to synchronize the action among different network devices. It is necessary to use the central control mode that described in [RFC8283] to correlate the forwarding behavior among different network devices. [RFC8821] describes the architecture and solution philosophy for the E2E traffic assurance in Native IP network via Multi Border Gateway Protocol (BGP) solution. This draft describes the corresponding Path Computation Element Communication Protocol (PCEP) extensions to transfer the key information about BGP peer info, peer prefix advertisement and the explicit peer route on on-path routers.

2. Conventions used in this document

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

3. Terminology

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

The following terms are defined in this document:

- o CCDD: Central Control Dynamic Routing
- o E2E: End to End
- o BPI: BGP Peer Info
- o EPR: Explicit Peer Route
- o PPA: Peer Prefix Advertisement
- o QoS: Quality of Service

4. Capability Advertisemnt

4.1. Open message

During the PCEP Initialization Phase, PCEP Speakers (PCE or PCC) advertise their support of Native IP extensions.

This document defines a new Path Setup Type (PST) [RFC8408] for Native-IP, as follows:

- o PST = TBD1: Path is a Native IP path as per [RFC8821].

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.

[I-D.ietf-pce-pcep-extension-for-pce-controller] defined the PCECC-CAPABILITY sub-TLV to exchange information about their PCECC capability. A new flag is defined in PCECC-CAPABILITY sub-TLV for Native IP:

N (NATIVE-IP-TE-CAPABILITY - 1 bit - TBD2): If set to 1 by a PCEP speaker, it indicates that the PCEP speaker is capable for TE in Native IP network 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 N 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 NATIVE-IP-TE-CAPABILITY bit is not set).
- o Terminate the PCEP session

5. PCEP messages

PCECC Native IP TE solution utilizing the existing PCE LSP Initiate Request message(PCInitiate) [RFC8281], and PCE Report message(PCRppt) [RFC8281] to accomplish the multi BGP sessions establishment, E2E TE path deployment, and route prefixes advertisement among different BGP sessions. A new PST for Native-IP is used to indicate the path setup based on TE in Native IP networks.

The extended PCInitiate message described in [I-D.ietf-pce-pcep-extension-for-pce-controller] is used to download or cleanup central controller's instructions (CCIs).

[I-D.ietf-pce-pcep-extension-for-pce-controller] specify an object called CCI for the encoding of central controller's instructions. This document specifies a new CCI object-type for Native IP. The PCEP messages are extended in this document to handle the PCECC operations for Native IP. Three new PCEP Objects (BGP Peer Info (BPI) Object, Explicit Peer Route (EPR) Object and Peer Prefix Advertisement (PPA) Object) are defined in this document. Refer to (Section 7) for detail object definitions.

5.1. The PCInitiate message

The PCInitiate Message defined in [RFC8281] and extended in [I-D.ietf-pce-pcep-extension-for-pce-controller] is further extended to support Native-IP CCI.

The format of the extended PCInitiate message is as follows:

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

Where:
  <Common Header> is defined in [RFC5440]

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

  <PCE-initiated-lsp-request> ::=
    (<PCE-initiated-lsp-instantiation> |
     <PCE-initiated-lsp-deletion> |
     <PCE-initiated-lsp-central-control>)

  <PCE-initiated-lsp-central-control> ::= <SRP>
                                          <LSP>
                                          (<cci-list> |
                                           ((<BPI> | <EPR> | <PPA>)
                                            <CCI>))

  <cci-list> ::= <CCI>
                [<cci-list>]
```

Where:

<cci-list> is as per
[I-D.ietf-pce-pcep-extension-for-pce-controller].
<PCE-initiated-lsp-instantiation> and
<PCE-initiated-lsp-deletion> are as per
[RFC8281].

The LSP and SRP objects are defined in [RFC8231].

When PCInitiate message is used create Native IP instructions, the SRP, LSP and CCI objects MUST be present. The error handling for missing SRP, LSP or CCI object is as per [I-D.ietf-pce-pcep-extension-for-pce-controller]. Further only one of BPI, EPR, or PPA object MUST be present. The PLSP-ID within the LSP object should be set by PCC uniquely according to the Symbolic Path Name TLV that included in the CCI object. The Symbolic Path Name is used by the PCE/PCC to identify uniquely the E2E native IP TE path.

If none of them are present, the receiving PCC MUST send a PCErr message with Error-type=6 (Mandatory Object missing) and Error-value=TBD4 (Native IP object missing). If there are more than one of BPI, EPR or PPA object are presented, the receiving PCC MUST send a PCErr message with Error-type=19 (Invalid Operation) and Error-value=TBD5 (Only one of the BPI, EPR or PPA object can be included in this message).

To cleanup the SRP object must set the R (remove) bit.

5.2. The PCRpt message

The PCRpt message is used to acknowledge the Native-IP instructions received from the central controller (PCE).

The format of the PCRpt message is as follows:

```

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

```

Where:

```

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

```

```

<state-report> ::= (<lsp-state-report>|
                    <central-control-report>)

```

```

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

```

```

<central-control-report> ::= [<SRP>]
                             <LSP>
                             (<cci-list>|
                              ((<BPI>|<EPR>|<PPA>)
                               <CCI>))

```

Where:

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

The error handling for missing CCI object is as per [I-D.ietf-pce-pcep-extension-for-pce-controller]. Further only one of BPI, EPR, or PPA object 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 (Native IP object missing). If there are more than one of BPI, EPR or PPA object are presented, the receiving PCE MUST send a PCErr message with Error-type=19(Invalid Operation) and Error-value=TBD5(Only one of the BPI, EPR or PPA object can be included in this message).

6. PCECC Native IP TE Procedures

The detail procedures for the TE in native IP environment are described in the following sections.

6.1. BGP Session Establishment Procedures

The procedures for establishing the BGP session between two peers is shown below, using the PCInitiate and PCRpt message pair.

The PCInitiate message should be sent to PCC which acts as BGP router and route reflector(RR). In the example in Figure 1, it should be sent to R1(M1), R3(M2 & M3) and R7(M4), when R3 acts as RR.

When PCC receives the BPI and CCI object (with the R bit set to 0 in SRP object) in PCInitiate message, the PCC should try to establish the BGP session with the indicated Peer AS and Local/Peer IP address.

When PCC creates successfully the BGP session that is indicated by the associated information, it should report the result via the PCRpt messages, with BPI object and the corresponding SRP and CCI object included.

When PCC receives this message with the R bit set to 1 in SRP object in PCInitiate message, the PCC should clear the BGP session that indicated by the BPI object.

When PCC clears successfully the specified BGP session, it should report the result via the PCRpt message, with the BPI object included, and the corresponding SRP and CCI object.

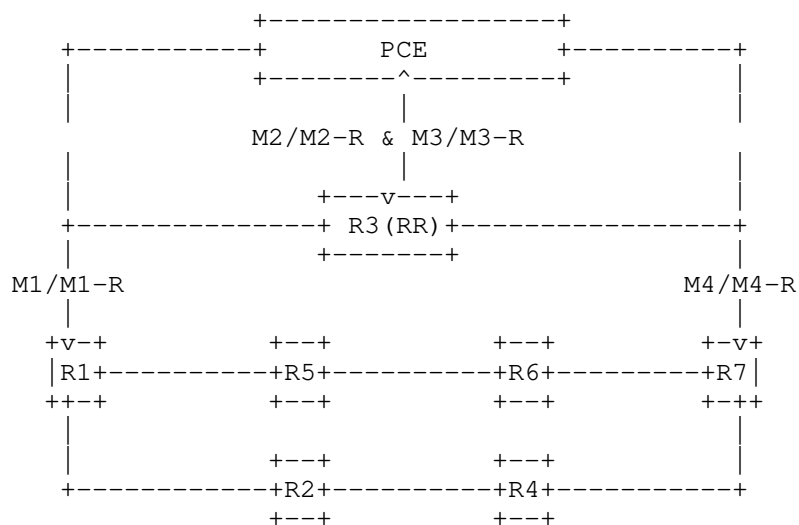


Figure 1: BGP Session Establishment Procedures(R3 act as RR)

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) BPI Object (Local_IP=R1_A, Peer_IP=R3_A)
M2 M2-R	PCE/R3	PCInitiate PCRpt	CC-ID=X2 (Symbolic Path Name=Class A) BPI Object (Local_IP=R3_A, Peer_IP=R1_A)
M3 M3-R	PCE/R3	PCInitiate PCRpt	CC-ID=X3 (Symbolic Path Name=Class A) BPI Object (Local_IP=R3_A, Peer_IP=R7_A)
M4 M4-R	PCE/R7	PCInitiate PCRpt	CC-ID=X4 (Symbolic Path Name=Class A) BPI Object (Local_IP=R7_A, Peer_IP=R3_A)

If the PCC cannot establish the BGP session that required by this object, it should report the error values via PCErr message with the newly defined error type (Error-type=TBD6) and error value (Error-value=TBD7, Peer AS not match; or Error-Value=TBD8, Peer IP can't be reached), which is indicated in Section 11

If the Local IP Address or Peer IP Address within BPI object is used in other existing BGP sessions, the PCC should report such error situation via PCErr message with Err-type=TBD6 and error value (Error-value=TBD9, Local IP is in use; Error-value=TBD10, Remote IP is in use).

6.2. Explicit Route Establish Procedures

The detail procedures for the explicit route establishment procedures is shown below, using PCInitiate and PCRpt message pair.

The PCInitiate message should be sent to the on-path routers respectively. In the example, for explicit route from R1 to R7, the PCInitiate message should be sent to R1(M1), R2(M2) and R4(M3), as shown in Figure 2. For explicit route from R7 to R1, the PCInitiate message should be sent to R7(M1), R4(M2) and R2(M3), as shown in Figure 3.

When PCC receives the EPR and the CCI object (with the R bit set to 0 in SRP object) in PCInitiate message, the PCC should install the explicit route to the peer.

When PCC install successfully the explicit route to the peer, it should report the result via the PCRpt messages, with EPR object and the corresponding SRP and CCI object included.

When PCC receives the EPR and the CCI object with the R bit set to 1 in SRP object in PCInitiate message, the PCC should clear the explicit route to the peer that indicated by the EPR object.

When PCC clear successfully the explicit route that indicated by this object, it should report the result via the PCRpt message, with the EPR object included, and the corresponding SRP and CCI object.

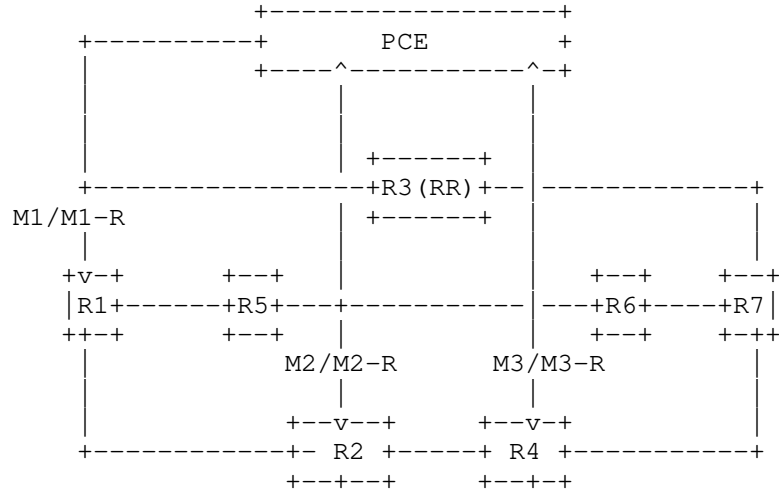


Figure 2: Explicit Route Establish Procedures (From R1 to R7)

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/R1	PCInitiate PCRpt	CC-ID=X1 (Symbolic Path Name=Class A) EPR Object (Peer Address=R7_A, Next Hop=R2_A)
M2 M2-R	PCE/R2	PCInitiate PCRpt	CC-ID=X2 (Symbolic Path Name=Class A) EPR Object (Peer Address=R7_A, Next Hop=R4_A)
M3 M3-R	PCE/R4	PCInitiate PCRpt	CC-ID=X3 (Symbolic Path Name=Class A) EPR Object (Peer Address=R7_A, Next Hop=R7_A)

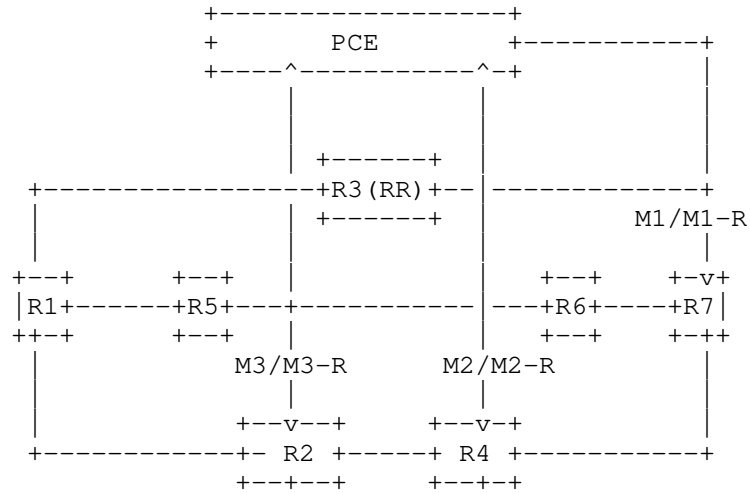


Figure 3: Explicit Route Establish Procedures (From R7 to R1)

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/R7	PCInitiate PCRpt	CC-ID=X1 (Symbolic Path Name=Class A) EPR Object (Peer Address=R1_A, Next Hop=R4_A)
M2 M2-R	PCE/R4	PCInitiate PCRpt	CC-ID=X2 (Symbolic Path Name=Class A) EPR Object (Peer Address=R1_A, Next Hop=R2_A)
M3 M3-R	PCE/R2	PCInitiate PCRpt	CC-ID=X3 (Symbolic Path Name=Class A) EPR Object (Peer Address=R1_A, Next Hop=R1_A)

In order to avoid the transient loop during the deploy of explicit peer route, the EPR object should be sent to the PCCs in the reverse order of the E2E path. To remove the explicit peer route, the EPR object should be sent to the PCCs in the same order of E2E path.

Upon the error occurs, the PCC SHOULD send the corresponding error via PCErr message, with an error information (Error-type=TBD6, Error-value=TBD12, Explicit Peer Route Error) that defined in Section 11.

When the peer info is not the same as the peer info that indicated in BPI object in PCC for the same path that is identified by Symbolic

Path Name TLV, an error (Error-type=TBD6, Error-value=17, EPR/BPI Peer Info mismatch) should be reported via the PCErr message.

6.3. BGP Prefix Advertisement Procedures

The detail procedures for BGP prefix advertisement are shown below, using PCInitiate and PCRpt message pair.

The PCInitiate message should be sent to PCC that acts as BGP peer router only. In the example, it should be sent to R1(M1) or R7(M2) respectively.

When PCC receives the PPA and the CCI object (with the R bit set to 0 in SRP object) in PCInitiate message, the PCC should send the prefixes indicated in this object to the appointed BGP peer.

When PCC sends successfully the prefixes to the appointed BGP peer, it should report the result via the PCRpt messages, with PPA object and the corresponding SRP and CCI object included.

When PCC receives the PPA and the CCI object with the R bit set to 1 in SRP object in PCInitiate message, the PCC should withdraw the prefixes advertisement to the peer that indicated by this object.

When PCC withdraws successfully the prefixes that indicated by this object, it should report the result via the PCRpt message, with the PPA object included, and the corresponding SRP and CCI object.

The IPv4 prefix MUST only be advertised via the IPv4 BGP session and the IPv6 prefix MUST only be advertised via the IPv6 BGP session. If mismatch occur, an error(Error-type=TBD6, Error-value=TBD18, BPI/PPR address family mismatch) should be reported via PCErr message.

When the peer info is not the same as the peer info that indicated in BPI object in PCC for the same path that is identified by Symbolic Path Name TLV, an error (Error-type=TBD6, Error-value=TBD19, PPA/BPI peer info mismatch) should be reported via the PCErr message.

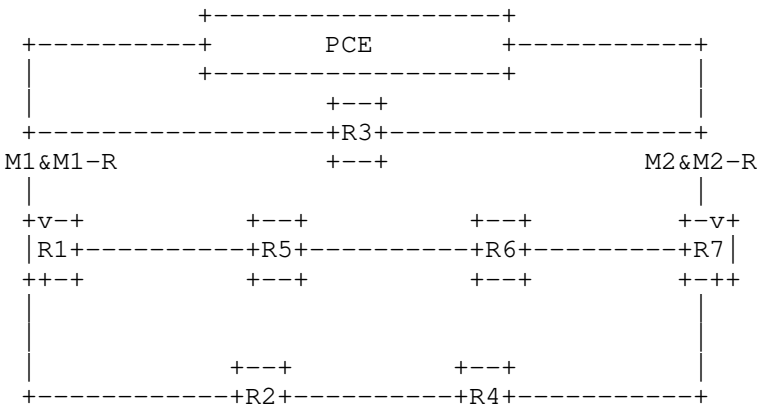


Figure 4: BGP Prefix Advertisement Procedures

Table 4: Message Information			
No.	Peers	Type	Message Key Parameters
M1 M1-R	PCE/R1	PCInitiate PCRpt	CC-ID=X1 (Symbolic Path Name=Class A) PPA Object (Peer IP=R7_A, Prefix=1_A)
M2 M2-R	PCE/R7	PCInitiate PCRpt	CC-ID=X2 (Symbolic Path Name=Class A) PPA Object (Peer IP=R1_A, Prefix=7_A)

7. New PCEP Objects

One new CCI Object and three new PCEP objects are defined in this draft. All new PCEP objects are as per [RFC5440]

7.1. CCI Object

The Central Control Instructions (CCI) Object is used by the PCE to specify the forwarding instructions is defined in [I-D.ietf-pce-pcep-extension-for-pce-controller]. This document defines another object-type for Native-IP.

CCI Object-Type is TBD13 for Native-IP as below

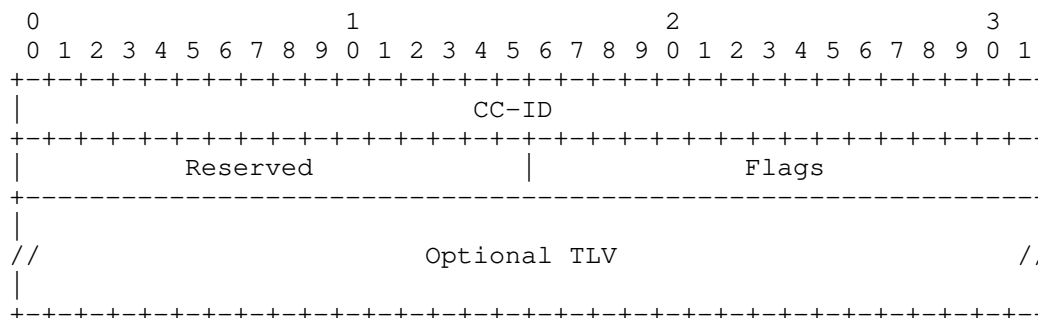


Figure 5: CCI Object for Native IP

Figure 1

The field CC-ID is as described in [I-D.ietf-pce-pcep-extension-for-pce-controller]. Following fields are defined for CCI Object-Type TBD13

Reserved: is set to zero while sending, ignored on receipt.

Flags: is used to carry any additional information pertaining to the CCI. Currently no flag bits are defined.

The Symbolic Path Name TLV [RFC8231] **MUST** be included in the CCI Object-Type TBD13 to identify the E2E TE path in Native IP environment and **MUST** be unique.

7.2. BGP Peer Info Object

The BGP Peer Info object is used to specify the information about the peer that the PCC should establish the BGP relationship with. This object should only be included and sent to the head and end router of the E2E path in case there is no Route Reflection (RR) involved. If the RR is used between the head and end routers, then such information should be sent to head router, RR and end router respectively.

By default, there **MUST** be no prefix be distributed via such BGP session that established by this object.

By default, the Local/Peer IP address **SHOULD** be dedicated to the usage of native IP TE solution, and **SHOULD NOT** be used by other BGP sessions that established by manual or non PCE initiated configuration.

BGP Peer Info Object-Class is TBD14

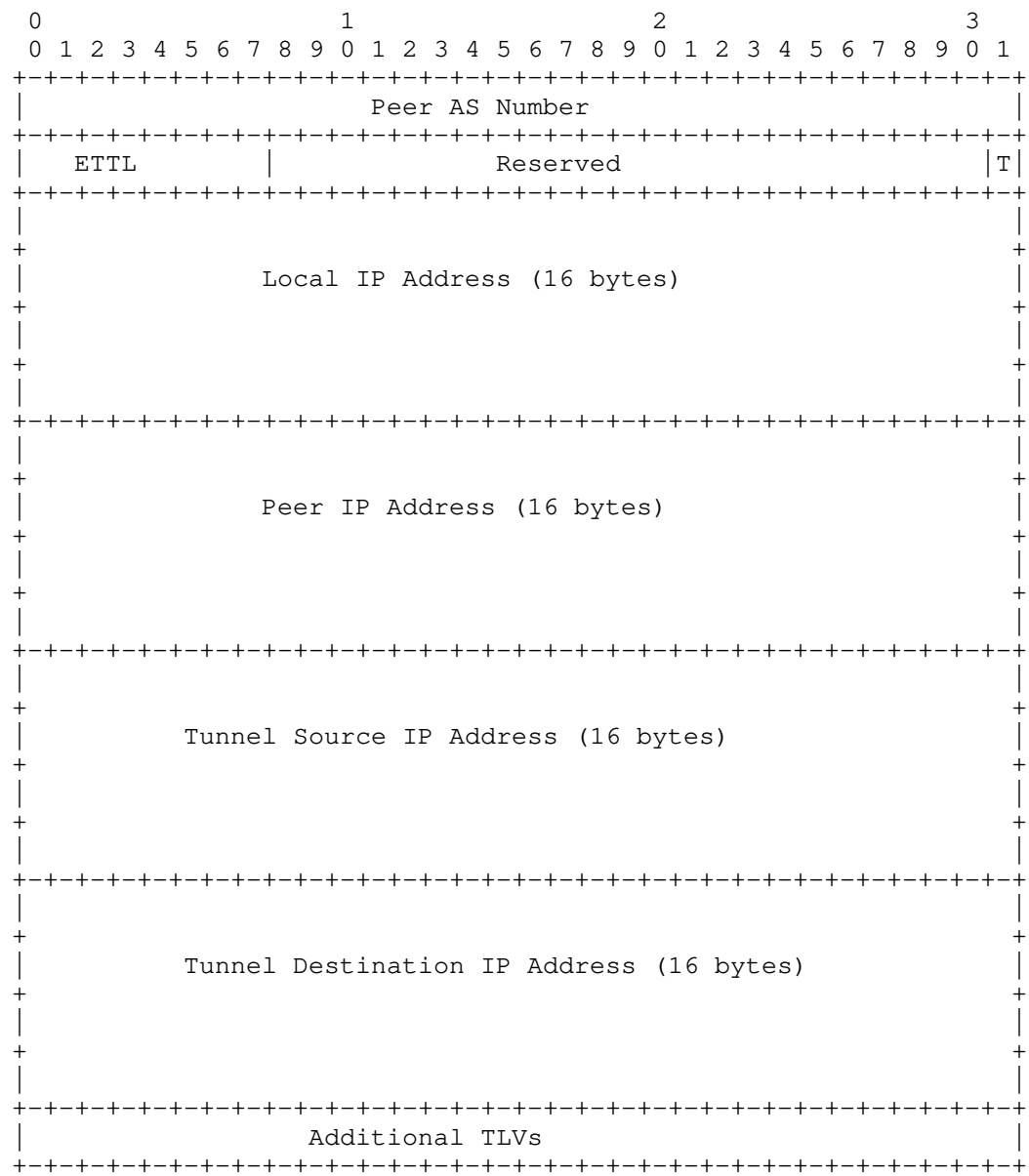


Figure 7: BGP Peer Info Object Body Format for IPv6

Peer AS Number: 4 Bytes, to indicate the AS number of Remote Peer.

ETTL: 1 Byte, to indicate the multi hop count for EBGp session. It should be 0 and ignored when Local AS and Peer AS is same.

Reserved: is set to zero while sending, ignored on receipt.

T bit: Indicates whether the traffic that associated with the prefixes advertised via this BGP session is transported via IPinIP tunnel (when T bit is set) or not (when T bit is clear).

Local IP Address(4/16 Bytes): IP address of the local router, used to peer with other end router. When Object-Type is 1, length is 4 bytes; when Object-Type is 2, length is 16 bytes.

Peer IP Address(4/16 Bytes): IP address of the peer router, used to peer with the local router. When Object-Type is 1, length is 4 bytes; when Object-Type is 2, length is 16 bytes;

Tunnel Source IP Address(4/16 Bytes): IP address of the tunnel source, should be owned by the local router. When Object-Type is 1, length is 4 bytes; when Object-Type is 2, length is 16 bytes.

Tunnel Destination IP Address(4/16 Bytes): IP address of the tunnel destination, should be owned by the peer router. When Object-Type is 1, length is 4 bytes; when Object-Type is 2, length is 16 bytes. Should be different from the Peer IP Address.

Additional TLVs: TLVs that associated with this object, can be used to convey other necessary information for dynamic BGP session establishment. Their definition are out of the current document.

When PCC receives BPI object, with Object-Type=1, it should try to establish BGP session with the peer in AFI/SAFI=1/1; when PCC receives BPI object with Object-Type=2, it should try to establish the BGP session with the peer in AFI/SAFI=2/1. Other BGP capabilities, for example, Graceful Restart (GR) that enhance the BGP performance should also be negotiated and used by default.

7.3. Explicit Peer Route Object

The Explicit Peer Route object is defined to specify the explicit peer route to the corresponding peer address on each device that is on the E2E assurance path. This Object should be sent to all the devices that locates on the E2E assurance path that calculated by PCE.

The path established by this object should have higher priority than other path calculated by dynamic IGP protocol, but should be lower priority than the static route configured by manual or NETCONF or by other means.

Explicit Peer Route Object-Class is TBD15.

Explicit Peer Route Object-Type is 1 for IPv4 and 2 for IPv6

The format of Explicit Peer Route object body for IPv4(Object-Type=1) is as follows:

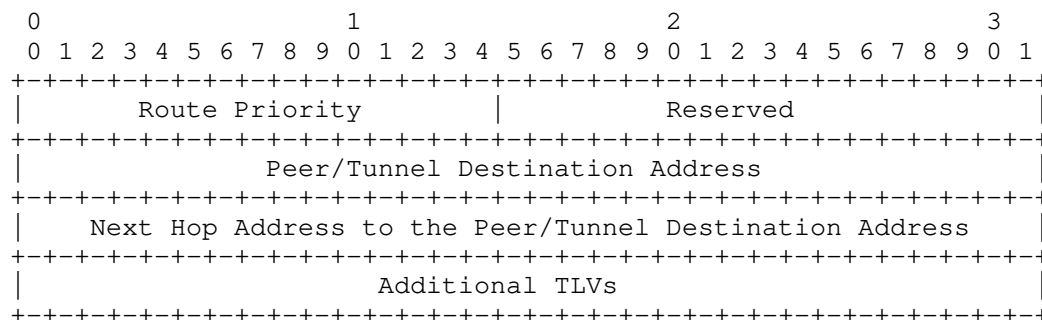


Figure 8: Explicit Peer Route Object Body Format for IPv4

The format of Explicit Peer Route object body for IPv6(Object-Type=2) is as follows:

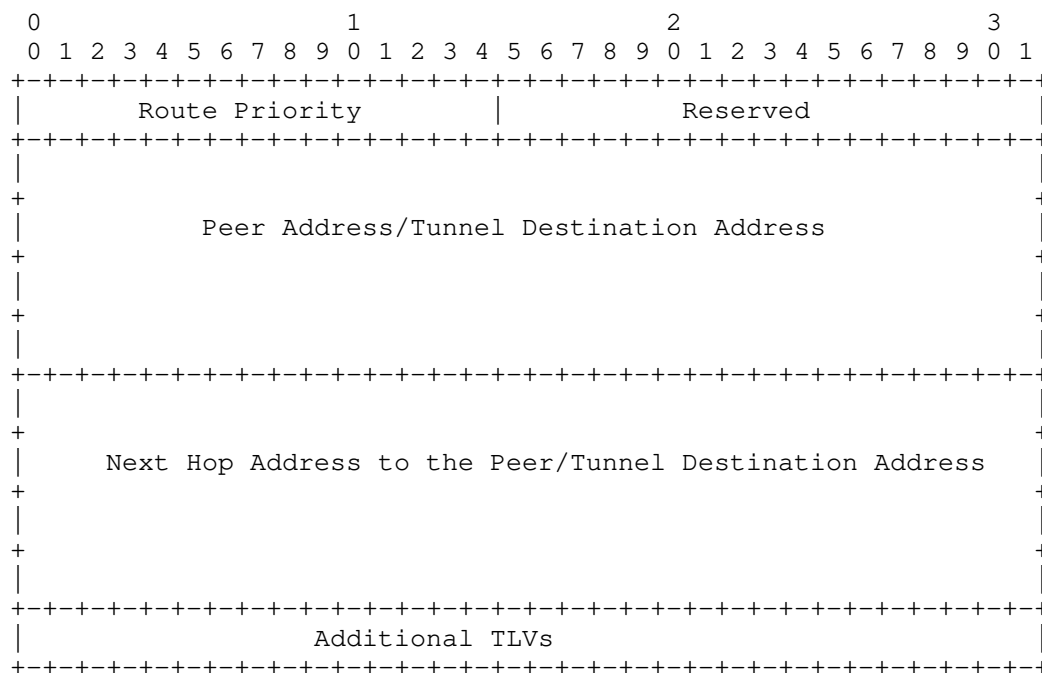


Figure 9: Explicit Peer Route Object Body Format for IPv6

Route Priority: 2 Bytes, The priority of this explicit route. The higher priority should be preferred by the device. This field is used to indicate the backup path at each hop.

Reserved.: is set to zero while sending, ignored on receipt.

Peer/Tunnel Destination Address: To indicate the peer address(4/16 Bytes). When T bit is set in the associated BPI object, use the tunnel destination address in BPI object; when T bit is clear, use the peer address in BPI object.

Next Hop Address to the Peer/Tunnel Destination Address: To indicate the next hop address(4/16 Bytes) to the corresponding peer/tunnel destination address.

Additional TLVs: TLVs that associated with this object, can be used to convey other necessary information for explicit peer path establishment. Its definition is out of the current document.

7.4. Peer Prefix Advertisement Object

The Peer Prefix Advertisement object is defined to specify the IP prefixes that should be advertised to the corresponding peer. This object should only be included and sent to the head/end router of the end2end path.

The prefixes information included in this object MUST only be advertised to the indicated peer, MUST NOT be advertised to other BGP peers.

Peer Prefix Advertisement Object-Class is TBD16

Peer Prefix Advertisement Object-Type is 1 for IPv4 and 2 for IPv6

The format of the Peer Prefix Advertisement object body is as follows:

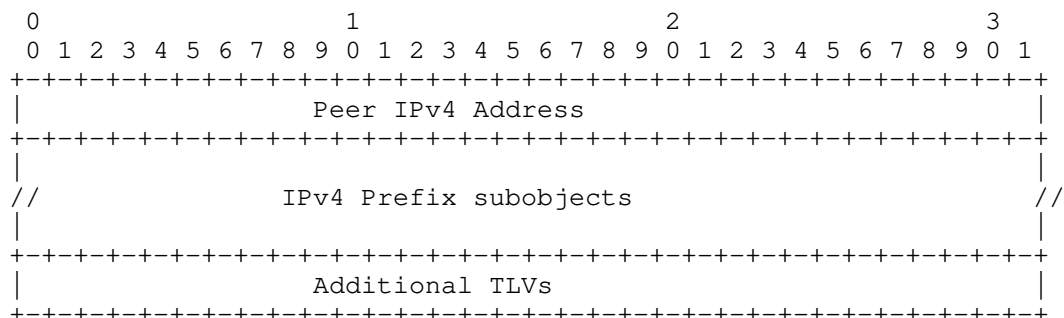


Figure 10: Peer Prefix Advertisement Object Body Format for IPv4

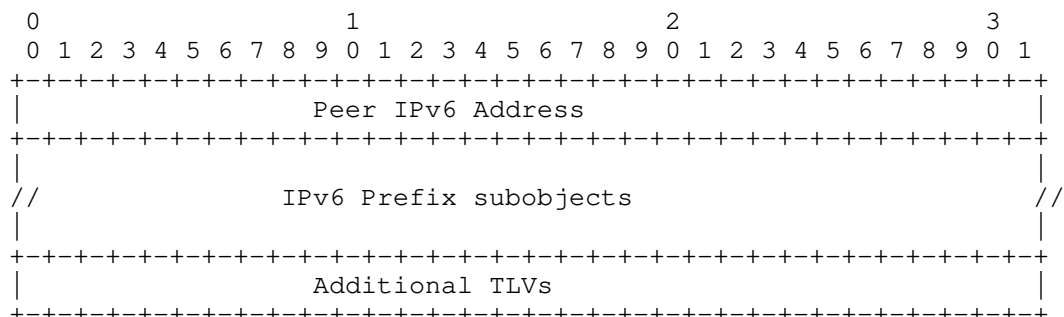


Figure 11: Peer Prefix Advertisement Object Body Format for IPv6

Peer IPv4 Address: 4 Bytes. Identifies the peer IPv4 address that the associated prefixes will be sent to.

IPv4 Prefix subobjects: List of IPv4 Prefix subobjects that defined in [RFC3209], identify the prefixes that will be sent to the peer that identified by Peer IPv4 Address List.

Peer IPv6 Address: 16 Bytes. Identifies the peer IPv6 address that the associated prefixes will be sent to.

IPv6 Prefix subobjects: List of IPv6 Prefix subobjects that defined in [RFC3209], identify the prefixes that will be sent to the peer that identified by Peer IPv6 Address List.

Additional TLVs: TLVs that associated with this object, can be used to convey other necessary information for prefixes advertisement. Its definition is out of the current document.

8. End to End Path Protection

[RFC8697] defines the path associations procedures between sets of Label Switched Path (LSP). Such procedures can also be used for the E2E path protection. To accomplish this, the PCE should attach the ASSOCIATION object with the EPR object in the PCInitiate message, with the association type set to 1 (Path Protection Association). The Extended Association ID that included within the Extended Association ID TLV, which is included in the ASSOCIATION object, should be set to the Symbolic Path Name of different E2E path. This PCInitiate should be sent to the head-end of the E2E path.

The head-end of the path can use the existing path detection mechanism, to monitor the status of the active path. Once it detects the failure, it can switch the backup protection path immediately.

9. Re-Delegation and Clean up

In case of a PCE failure, a new PCE can gain control over the central controller instructions. As per the PCEP procedures in [RFC8281], the State Timeout Interval timer is used to ensure that a PCE failure does not result in automatic and immediate disruption for the services. Similarly, as per [I-D.ietf-pcep-pcep-extension-for-pce-controller], the central controller instructions are not removed immediately upon PCE failure. Instead, they could be re-delegated to the new PCE before the expiration of this timer, or be cleaned up on the expiration of this timer. This allows for network clean up without manual intervention. The PCC MUST support the removal of CCI as one of the behaviors applied on expiration of the State Timeout Interval timer.

10. BGP Considerations

This draft defines the procedures and objects to create the BGP sessions and advertises the associated prefixes dynamically. Only the key information, for example peer IP addresses, peer AS number are exchanged via the PCEP protocol. Other parameters that are needed for the BGP session setup should be derived from their default values, as described in Section 7.2. Upon receiving such key information, the BGP module on the PCC should try to accomplish the task that appointed by the PCEP protocol and report the status to the PCEP modules.

There is no influence to current implementation of BGP Finite State Machine(FSM). The PCEP cares only the success and failure status of BGP session, and act upon such information accordingly.

The error handling procedures related to incorrect BGP parameters are specified in Section 6.1, Section 6.2, and Section 6.3. The handling of the dynamic BGP sessions and associated prefixes on PCE failure is described in Section 9.

11. New Error-Types and Error-Values Defined

A PCEP-ERROR object is used to report a PCEP error and is characterized by an Error-Type that specifies that type of error and an Error-value that provides additional information about the error. An additional Error-Type and several Error-values are defined to represent some the errors related to the newly defined objects, which are related to Native IP TE procedures.

Error-Type	Meaning	Error-value
TBD6	Native IP TE failure	
		0: Unassigned
		TBD7: Peer AS not match
		TBD8:Peer IP can't be reached
		TBD9:Local IP is in use
		TBD10:Remote IP is in use
		TBD11:Exist BGP session broken
		TBD12:Explicit Peer Route Error
		TBD17:EPR/BPI Peer Info mismatch
		TBD18:BPI/PPA Address Family mismatch
		TBD19:PPA/BPI Peer Info mismatch

Figure 12: Newly defined Error-Type and Error-Value

12. Deployment Considerations

The information transferred in this draft is mainly used for the light weight BGP session setup, explicit route deployment and the prefix distribution. The planning, allocation and distribution of

the peer addresses within IGP should be accomplished in advanced and they are out of the scope of this draft.

[RFC8232] describes the state synchronization procedure between stateful PCE and PCC. The communication of PCE and PCC described in this draft should also follow this procedures, treat the three newly defined objects that associated with the same symbolic path name as the attribute of the same path in the LSP-DB.

When PCE detects one or some of the PCCs are out of control, it should recompute and redeploy the traffic engineering path for native IP on the active PCCs. When PCC detects that it is out of control of the PCE, it should clear the information that initiated by the PCE. The PCE should assures the avoidance of possible transient loop in such node failure when it deploy the explicit peer route on the PCCs.

If the established BGP session is broken after some time, the PCC should also report such error via PCErr message with Err-type=TBD6 and error value(Error-value=TBD11, Existing BGP session is broken). Upon receiving such PCErr message, the PCE should clear the prefixes advertisement on the previous BGP session, clear the explicit peer route to the previous peer address; select other Local_IP/Peer_IP pair to establish the new BGP session, deploy the explicit peer route to the new peer address, and advertises the prefixes on the new BGP session.

13. Security Considerations

The setup of BGP sessions, prefix advertisement, and explicit peer route establishment are all controlled by the PCE. See [RFC4271] and [RFC4272] for BGP security considerations. Security consideration part in [RFC5440] and [RFC8231] should be considered. To prevent a bogus PCE sending harmful messages to the network nodes, the network devices should authenticate the validity of the PCE and ensure a secure communication channel between them. Mechanisms described in [RFC8253] should be used.

14. IANA Considerations

14.1. 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	Native IP TE Path	This document

14.2. PCECC-CAPABILITY sub-TLV's Flag field

[I-D.ietf-pce-pcep-extension-for-pce-controller] 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 (N)	NATIVE-IP-TE-CAPABILITY	This document

14.3. PCEP Object Types

IANA is requested to allocate new registry for the PCEP Object Type:

Object-Class Value	Name	Reference
44	CCI Object Object-Type TBD13: Native IP	This document
TBD14	BGP Peer Info Object-Type 1: IPv4 address 2: IPv6 address	This document
TBD15	Explicit Peer Route Object-Type 1: IPv4 address 2: IPv6 address	This document
TBD16	Peer Prefix Advertisement Object-Type 1: IPv4 address 2: IPv6 address	This document

14.4. 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:Native IP object missing This document
10	Reception of an invalid object	TBD3:PCECC NATIVE-IP-TE-CAPABILITY bit is not set This document
19	Invalid Operation	TBD5:Only one of the BPI,EPR or PPA object can be included in this message This document
TBD6	Native IP TE failure	This document TBD7:Peer AS not match TBD8:Peer IP can't be reached TBD9:Local IP is in use TBD10:Remote IP is in use TBD11:Exist BGP session broken TBD12:Explicit Peer Route Error TBD17:EPR/BPI Peer Info mismatch TBD18:BPI/PPA Address Family mismatch TBD19:PPA/BPI Peer Info mismatch

15. Contributor

Dhruv Dhody has contributed the contents of this draft.

16. Acknowledgement

Thanks Mike Koldychev, Siva Sivabalan, Adam Simpson for his valuable suggestions and comments.

17. Normative References

- [I-D.ietf-pce-pcep-extension-for-pce-controller]
Li, Z., Peng, S., Negi, M. S., Zhao, Q., and C. Zhou,
"PCEP Procedures and Protocol Extensions for Using PCE as
a Central Controller (PCECC) of LSPs", draft-ietf-pce-
pcep-extension-for-pce-controller-14 (work in progress),
March 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>>.
- [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>>.

- [RFC4271] Rekhter, Y., Ed., Li, T., Ed., and S. Hares, Ed., "A Border Gateway Protocol 4 (BGP-4)", RFC 4271, DOI 10.17487/RFC4271, January 2006, <<https://www.rfc-editor.org/info/rfc4271>>.
- [RFC4272] Murphy, S., "BGP Security Vulnerabilities Analysis", RFC 4272, DOI 10.17487/RFC4272, January 2006, <<https://www.rfc-editor.org/info/rfc4272>>.
- [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>>.
- [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>>.
- [RFC8253] Lopez, D., Gonzalez de Dios, O., Wu, Q., and D. Dhody, "PCEPS: Usage of TLS to Provide a Secure Transport for the Path Computation Element Communication Protocol (PCEP)", RFC 8253, DOI 10.17487/RFC8253, October 2017, <<https://www.rfc-editor.org/info/rfc8253>>.
- [RFC8281] Crabbe, E., Minei, I., Sivabalan, S., and R. Varga, "Path Computation Element Communication Protocol (PCEP) Extensions for PCE-Initiated LSP Setup in a Stateful PCE Model", RFC 8281, DOI 10.17487/RFC8281, December 2017, <<https://www.rfc-editor.org/info/rfc8281>>.
- [RFC8283] Farrel, A., Ed., Zhao, Q., Ed., Li, Z., and C. Zhou, "An Architecture for Use of PCE and the PCE Communication Protocol (PCEP) in a Network with Central Control", RFC 8283, DOI 10.17487/RFC8283, December 2017, <<https://www.rfc-editor.org/info/rfc8283>>.

- [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>>.
- [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>>.
- [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>>.
- [RFC8821] Wang, A., Khasanov, B., Zhao, Q., and H. Chen, "PCE-Based Traffic Engineering (TE) in Native IP Networks", RFC 8821, DOI 10.17487/RFC8821, April 2021, <<https://www.rfc-editor.org/info/rfc8821>>.

Authors' Addresses

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
Russia

Email: bhassanov@yahoo.com

Sheng Fang
Huawei Technologies, Co., Ltd
Huawei Bld., No.156 Beiqing Rd.
Beijing
China

Email: fsheng@huawei.com

Ren Tan
Huawei Technologies, Co., Ltd
Huawei Bld., No.156 Beiqing Rd.
Beijing
China

Email: tanren@huawei.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: 22 September 2022

A. Wang
China Telecom
B. Khasanov
Yandex LLC
S. Fang
R. Tan
Huawei Technologies, Co., Ltd
C. Zhu
ZTE Corporation
21 March 2022

PCEP Extension for Native IP Network
draft-ietf-pce-pcep-extension-native-ip-18

Abstract

This document defines the Path Computation Element Communication Protocol (PCEP) extension for Central Control Dynamic Routing (CCDR) based application in Native IP network. The scenario and framework of CCDR in native IP is described in [RFC8735] and [RFC8821]. This draft describes the key information that is transferred between Path Computation Element (PCE) and Path Computation Clients (PCC) to accomplish the End to End (E2E) traffic assurance in Native IP network under central control mode.

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 22 September 2022.

Copyright Notice

Copyright (c) 2022 IETF Trust and the persons identified as the document authors. All rights reserved.

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

Table of Contents

1. Introduction	3
2. Conventions used in this document	3
3. Terminology	3
4. Capability Advertisemnt	4
4.1. Open message	4
5. PCEP messages	4
5.1. The PCInitiate message	5
5.2. The PCRpt message	6
6. PCECC Native IP TE Procedures	7
6.1. BGP Session Establishment Procedures	7
6.2. Explicit Route Establish Procedures	9
6.3. BGP Prefix Advertisement Procedures	12
7. New PCEP Objects	14
7.1. CCI Object	14
7.2. BGP Peer Info Object	15
7.3. Explicit Peer Route Object	17
7.4. Peer Prefix Advertisement Object	20
8. End to End Path Protection	21
9. Re-Delegation and Clean up	21
10. BGP Considerations	22
11. New Error-Types and Error-Values Defined	22
12. Deployment Considerations	23
13. Implementation Status	24
13.1. Proof of Concept based on ODL	24
14. Security Considerations	25
15. IANA Considerations	25
15.1. Path Setup Type Registry	25
15.2. PCECC-CAPABILITY sub-TLV's Flag field	25
15.3. PCEP Object Types	25
15.4. PCEP-Error Object	26
16. Contributor	27
17. Acknowledgement	27
18. Normative References	27
Authors' Addresses	29

1. Introduction

Generally, Multiprotocol Label Switching Traffic Engineering (MPLS-TE) requires the corresponding network devices support Multiprotocol Label Switching (MPLS) or Resource ReSerVation Protocol (RSVP)/Label Distribution Protocol (LDP) technologies to assure the End-to-End (E2E) traffic performance. In Segment Routing either IGP extensions or BGP are used to steer a packet through an SR Policy instantiated as an ordered list of instructions called "segments". But in native IP network, there will be no such signaling protocol to synchronize the action among different network devices. It is necessary to use the central control mode that described in [RFC8283] to correlate the forwarding behavior among different network devices. [RFC8821] describes the architecture and solution philosophy for the E2E traffic assurance in Native IP network via Multi Border Gateway Protocol (BGP) solution. This draft describes the corresponding Path Computation Element Communication Protocol (PCEP) extensions to transfer the key information about BGP peer info, peer prefix advertisement and the explicit peer route on on-path routers.

2. Conventions used in this document

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

3. Terminology

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

The following terms are defined in this document:

- * CCDR: Central Control Dynamic Routing
- * E2E: End to End
- * BPI: BGP Peer Info
- * EPR: Explicit Peer Route
- * PPA: Peer Prefix Advertisement
- * QoS: Quality of Service

4. Capability Advertisemnt

4.1. Open message

During the PCEP Initialization Phase, PCEP Speakers (PCE or PCC) advertise their support of Native IP extensions.

This document defines a new Path Setup Type (PST) [RFC8408] for Native-IP, as follows:

- * PST = TBD1: Path is a Native IP path as per [RFC8821].

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.

[RFC9050] defined the PCECC-CAPABILITY sub-TLV to exchange information about their PCECC capability. A new flag is defined in PCECC-CAPABILITY sub-TLV for Native IP:

N (NATIVE-IP-TE-CAPABILITY - 1 bit - TBD2): If set to 1 by a PCEP speaker, it indicates that the PCEP speaker is capable for TE in Native IP network 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 N 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 NATIVE-IP-TE-CAPABILITY bit is not set).
- * Terminate the PCEP session

5. PCEP messages

PCECC Native IP TE solution utilizing the existing PCE LSP Initiate Request message(PCInitiate)[RFC8281], and PCE Report message(PCRppt)[RFC8281] to accomplish the multi BGP sessions establishment, E2E TE path deployment, and route prefixes advertisement among different BGP sessions. A new PST for Native-IP is used to indicate the path setup based on TE in Native IP networks.

The extended PCInitiate message described in [RFC9050] is used to download or cleanup central controller's instructions (CCIs). [RFC9050] specifies an object called CCI for the encoding of central controller's instructions. This document specify a new CCI object-

type for Native IP. The PCEP messages are extended in this document to handle the PCECC operations for Native IP. Three new PCEP Objects (BGP Peer Info (BPI) Object, Explicit Peer Route (EPR) Object and Peer Prefix Advertisement (PPA) Object) are defined in this document. Refer to Section 7 for detail object definitions.

5.1. The PCInitiate message

The PCInitiate Message defined in [RFC8281] and extended in [RFC9050] is further extended to support Native-IP CCI.

The format of the extended PCInitiate message is as follows:

```

<PCInitiate Message> ::= <Common Header>
                           <PCE-initiated-lsp-list>
Where:
  <Common Header> is defined in [RFC5440]

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

  <PCE-initiated-lsp-request> ::=
    (<PCE-initiated-lsp-instantiation>|
     <PCE-initiated-lsp-deletion>|
     <PCE-initiated-lsp-central-control>)

  <PCE-initiated-lsp-central-control> ::= <SRP>
                                          <LSP>
                                          (<cci-list>|
                                           ((<BPI>|<EPR>|<PPA>)
                                            <CCI>))

  <cci-list> ::= <CCI>
                [<cci-list>]

```

Where:

<cci-list> is as per
 [I-D.ietf-pce-pcep-extension-for-pce-controller].
 <PCE-initiated-lsp-instantiation> and
 <PCE-initiated-lsp-deletion> are as per
 [RFC8281].

The LSP and SRP objects are defined in [RFC8231].

When PCInitiate message is used create Native IP 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, EPR, or PPA object MUST be present. The PLSP-ID within the

LSP object should be set by PCC uniquely according to the Symbolic Path Name TLV that included in the CCI object. The Symbolic Path Name is used by the PCE/PCC to identify uniquely the E2E native IP TE path.

If none of them are present, the receiving PCC MUST send a PCErr message with Error-type=6 (Mandatory Object missing) and Error-value=TBD4 (Native IP object missing). If there are more than one of BPI, EPR or PPA object are presented, the receiving PCC MUST send a PCErr message with Error-type=19 (Invalid Operation) and Error-value=TBD5 (Only one of the BPI, EPR or PPA object can be included in this message).

To cleanup the SRP object must set the R (remove) bit.

5.2. The PCRpt message

The PCRpt message is used to acknowledge the Native-IP instructions received from the central controller (PCE).

The format of the PCRpt message is as follows:

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

Where:

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

```
<state-report> ::= (<lsp-state-report>|
                    <central-control-report>)
```

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

```
<central-control-report> ::= [<SRP>]
                              <LSP>
                              (<cci-list>|
                               ((<BPI>|<EPR>|<PPA>)
                                <CCI>))
```

Where:

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

The error handling for missing CCI object is as per [RFC9050]. Further only one of BPI, EPR, or PPA object 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 (Native IP object missing). If there are more than one of BPI, EPR or PPA object are presented, the receiving PCE MUST send a PCErr message with Error-type=19(Invalid Operation) and Error-value=TBD5(Only one of the BPI, EPR or PPA object can be included in this message).

6. PCECC Native IP TE Procedures

The detail procedures for the TE in native IP environment are described in the following sections.

6.1. BGP Session Establishment Procedures

The PCInitiate message can be used to configure the parameters for a BGP peer session using the PCInitiate and PCRpt message pair. This pair of PCE messages is exchanged with a PCE function attached to each BGP peer which needs to be configured. After the BGP peer session has been configured via this pair of PCE messages the BGP session establishment process operates in a normal fashion. All BGP peers are configured for peer to peer communication whether the peers are E-BGP peers or I-BGP peers. One of the IBGP topologies requires that multiple I-BGPs peers operate in a route-reflector I-BGP peer topology. The example below shows two I-BGP route reflector clients interacting with one Route Reflector (RR), but Route Reflector topologies may have up to 100s of clients. Centralized configuration via PCE provides mechanisms to scale auto-configuration of small and large topologies.

The PCInitiate message should be sent to PCC which acts as BGP router and/or route reflector(RR).

The route reflector topology for a single AS is shown in Figure 1. The BGP routers R1, R3, and R7 are within a single AS. R1 and R7 are BGP router-reflector clients, and R3 is a Route Reflector. The PCInitiate message should be sent all of the BGP routers that need to be configured R1 (M3), R3 (M2 & M3), and R7 (M4).

PCInitiate message creates an auto-configuration function for these BGP peers providing the indicated Peer AS and the Local/Peer IP Address.

When PCC receives the BPI and CCI object (with the R bit set to 0 in SRP object) in PCInitiate message, the PCC should try to establish the BGP session with the indicated Peer AS and Local/Peer IP address.

When PCC creates successfully the BGP session that is indicated by the associated information, it should report the result via the PCRpt messages, with BPI object and the corresponding SRP and CCI object included.

When PCC receives this message with the R bit set to 1 in SRP object in PCInitiate message, the PCC should clear the BGP session that indicated by the BPI object.

When PCC clears successfully the specified BGP session, it should report the result via the PCRpt message, with the BPI object included, and the corresponding SRP and CCI object.

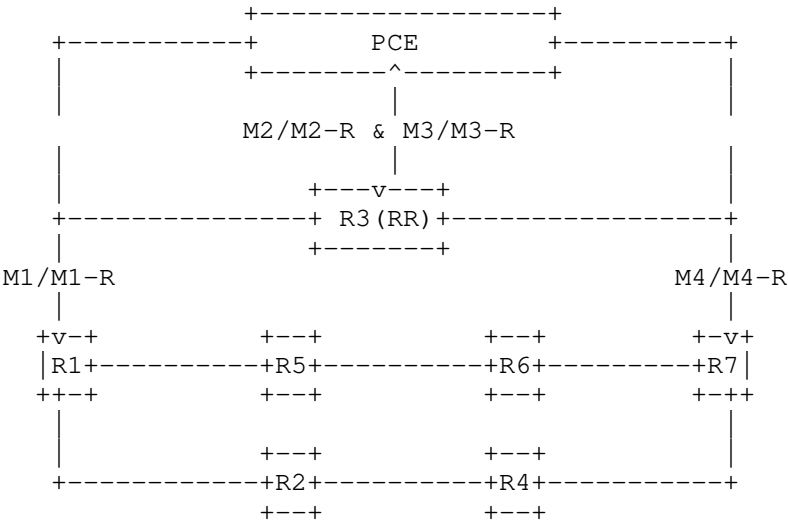


Figure 1: BGP Session Establishment Procedures(R3 act as RR)

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) BPI Object (Local_IP=R1_A, Peer_IP=R3_A)
M2 M2-R	PCE/R3	PCInitiate PCRpt	CC-ID=X2 (Symbolic Path Name=Class A) BPI Object (Local_IP=R3_A, Peer_IP=R1_A)
M3 M3-R	PCE/R3	PCInitiate PCRpt	CC-ID=X3 (Symbolic Path Name=Class A) BPI Object (Local_IP=R3_A, Peer_IP=R7_A)
M4 M4-R	PCE/R7	PCInitiate PCRpt	CC-ID=X4 (Symbolic Path Name=Class A) BPI Object (Local_IP=R7_A, Peer_IP=R3_A)

If the PCC cannot establish the BGP session that required by this object, it should report the error values via PCErr message with the newly defined error type (Error-type=TBD6) and error value (Error-value=TBD7, Peer AS not match; or Error-Value=TBD8, Peer IP can't be reached), which is indicated in Section 11

If the Local IP Address or Peer IP Address within BPI object is used in other existing BGP sessions, the PCC should report such error situation via PCErr message with Err-type=TBD6 and error value (Error-value=TBD9, Local IP is in use; Error-value=TBD10, Remote IP is in use).

6.2. Explicit Route Establish Procedures

The explicit route establishment procedures can be used to install a route via PCE in the PCC/BGP Peer, using PCInitiate and PCRpt message pair. Although the BGP policy might redistribute the routes installed by explicit route, the PCE-BGP implementation needs to prohibit the redistribution of the explicit route. PCE explicit routes operate similar to static routes installed by network management protocols (netconf/restconf) but the routes are associated with the PCE routing module. Explicit route installations (like NM static routes) must carefully install and uninstall static routes in an specific order so that the pathways are established without loops.

The PCInitiate message should be sent to the on-path routers respectively. In the example, for explicit route from R1 to R7, the PCInitiate message should be sent to R1 (M1), R2 (M2) and R4 (M3), as shown in Figure 2. For explicit route from R7 to R1, the PCInitiate message should be sent to R7 (M1), R4 (M2) and R2 (M3), as shown in Figure 3.

When PCC receives the EPR and the CCI object (with the R bit set to 0 in SRP object) in PCInitiate message, the PCC should install the explicit route to the the peer.

When PCC install successfully the explicit route to the peer, it should report the result via the PCRpt messages, with EPR object and the corresponding SRP and CCI object included.

When PCC receives the EPR and the CCI object with the R bit set to 1 in SRP object in PCInitiate message, the PCC should clear the explicit route to the peer that indicated by the EPR object.

When PCC clear successfully the explicit route that indicated by this object, it should report the result via the PCRpt message, with the EPR object included, and the corresponding SRP and CCI object.

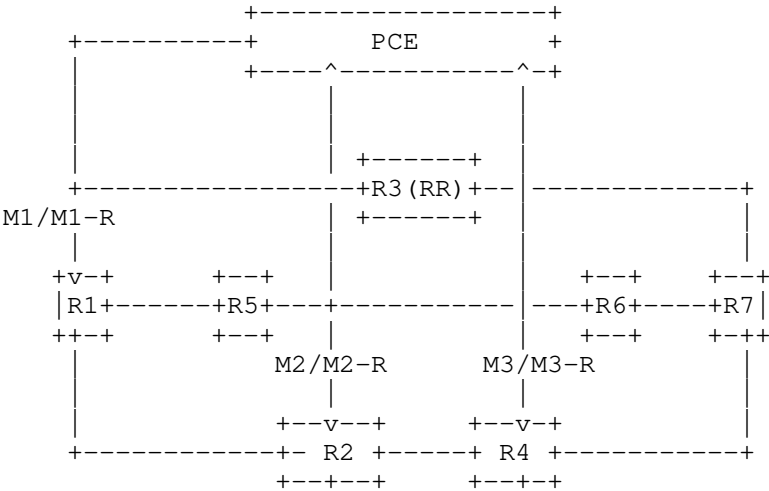


Figure 2: Explicit Route Establish Procedures (From R1 to R7)

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/R1	PCInitiate PCRpt	CC-ID=X1 (Symbolic Path Name=Class A) EPR Object (Peer Address=R7_A, Next Hop=R2_A)
M2 M2-R	PCE/R2	PCInitiate PCRpt	CC-ID=X2 (Symbolic Path Name=Class A) EPR Object (Peer Address=R7_A, Next Hop=R4_A)
M3 M3-R	PCE/R4	PCInitiate PCRpt	CC-ID=X3 (Symbolic Path Name=Class A) EPR Object (Peer Address=R7_A, Next Hop=R7_A)

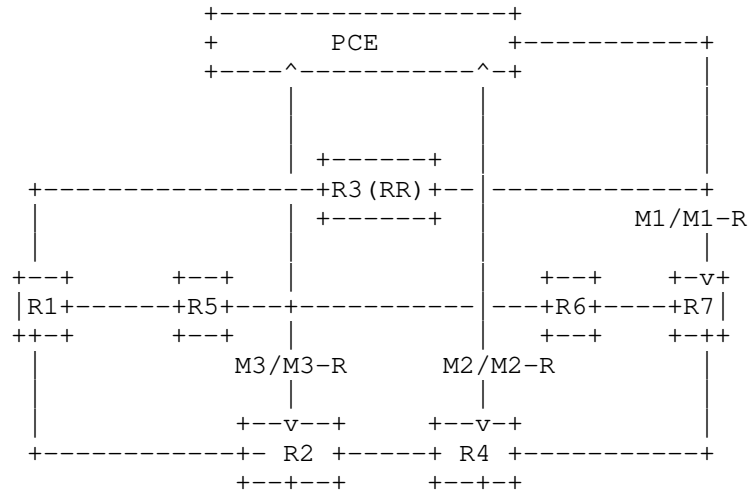


Figure 3: Explicit Route Establish Procedures (From R7 to R1)

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/R7	PCInitiate PCRpt	CC-ID=X1 (Symbolic Path Name=Class A) EPR Object (Peer Address=R1_A, Next Hop=R4_A)
M2 M2-R	PCE/R4	PCInitiate PCRpt	CC-ID=X2 (Symbolic Path Name=Class A) EPR Object (Peer Address=R1_A, Next Hop=R2_A)
M3 M3-R	PCE/R2	PCInitiate PCRpt	CC-ID=X3 (Symbolic Path Name=Class A) EPR Object (Peer Address=R1_A, Next Hop=R1_A)

In order to avoid the transient loop during the deploy of explicit peer route, the EPR object should be sent to the PCCs in the reverse order of the E2E path. To remove the explicit peer route, the EPR object should be sent to the PCCs in the same order of E2E path.

To accomplish ECMP effects, the PCE can send multiple EPR objects to the same node, with the same route priority and peer address value but different next hop addresses.

The PCC should verify that the next hop address is reachable. Upon the error occurs, the PCC SHOULD send the corresponding error via PCErr message, with an error information (Error-type=TBD6, Error-value=TBD12, Explicit Peer Route Error) that defined in Section 11.

When the peer info is not the same as the peer info that indicated in BPI object in PCC for the same path that is identified by Symbolic Path Name TLV, an error (Error-type=TBD6, Error-value=17, EPR/BPI Peer Info mismatch) should be reported via the PCErr message.

6.3. BGP Prefix Advertisement Procedures

The detail procedures for BGP prefix advertisement are shown below, using PCInitiate and PCRpt message pair.

The PCInitiate message should be sent to PCC that acts as BGP peer router only. In the example, it should be sent to R1(M1) or R7(M2) respectively.

When PCC receives the PPA and the CCI object (with the R bit set to 0 in SRP object) in PCInitiate message, the PCC should send the prefixes indicated in this object to the appointed BGP peer.

When PCC sends successfully the prefixes to the appointed BGP peer, it should report the result via the PCRpt messages, with PPA object and the corresponding SRP and CCI object included.

When PCC receives the PPA and the CCI object with the R bit set to 1 in SRP object in PCInitiate message, the PCC should withdraw the prefixes advertisement to the peer that indicated by this object.

When PCC withdraws successfully the prefixes that indicated by this object, it should report the result via the PCRpt message, with the PPA object included, and the corresponding SRP and CCI object.

The allowed AFI/SAFI for the IPv4 BGP session should be 1/1(IPv4 prefix) and the allowed AFI/SAFI for the IPv6 BGP session should be 2/1(IPv6 prefix). If mismatch occur, an error(Error-type=TBD6, Error-value=TBD18, BPI/PPR address family mismatch) should be reported via PCErr message.

When the peer info is not the same as the peer info that indicated in BPI object in PCC for the same path that is identified by Symbolic Path Name TLV, an error (Error-type=TBD6, Error-value=TBD19, PPA/BPI peer info mismatch) should be reported via the PCErr message.

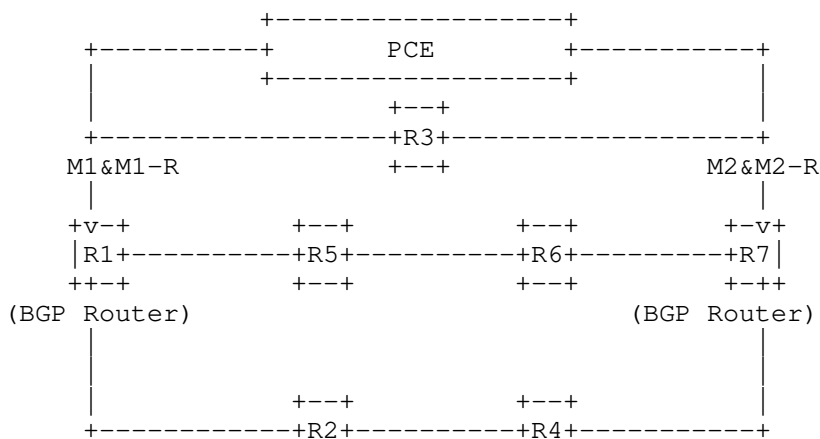


Figure 4: BGP Prefix Advertisement Procedures

Table 4: Message Information

No.	Peers	Type	Message Key Parameters
M1 M1-R	PCE/R1	PCInitiate PCRpt	CC-ID=X1 (Symbolic Path Name=Class A) PPA Object (Peer IP=R7_A, Prefix=1_A)
M2 M2-R	PCE/R7	PCInitiate PCRpt	CC-ID=X2 (Symbolic Path Name=Class A) PPA Object (Peer IP=R1_A, Prefix=7_A)

7. New PCEP Objects

One new CCI Object and three new PCEP objects are defined in this draft. All new PCEP objects are as per [RFC5440]

7.1. 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 Native-IP.

CCI Object-Type is TBD13 for Native-IP as below

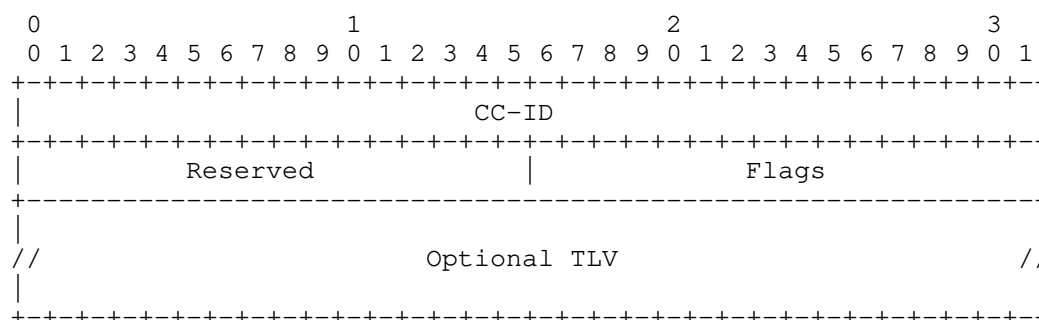


Figure 5: CCI Object for Native IP

Figure 1

The field CC-ID is as described in [RFC9050]. Following fields are defined for CCI Object-Type TBD13

Reserved: is set to zero while sending, ignored on receipt.

Flags: is used to carry any additional information pertaining to the CCI. Currently no flag bits are defined.

The Symbolic Path Name TLV [RFC8231] MUST be included in the CCI Object-Type TBD13 to identify the E2E TE path in Native IP environment and MUST be unique.

7.2. BGP Peer Info Object

The BGP Peer Info object is used to specify the information about the peer that the PCC should establish the BGP relationship with. This object should only be included and sent to the head and end router of the E2E path in case there is no Route Reflection (RR) involved. If the RR is used between the head and end routers, then such information should be sent to head router, RR and end router respectively.

By default, there MUST be no prefix be distributed via such BGP session that established by this object.

By default, the Local/Peer IP address SHOULD be dedicated to the usage of native IP TE solution, and SHOULD NOT be used by other BGP sessions that established by manual or non PCE initiated configuration.

BGP Peer Info Object-Class is TBD14

BGP Peer Info Object-Type is 1 for IPv4 and 2 for IPv6

The format of the BGP Peer Info object body for IPv4 (Object-Type=1) is as follows:

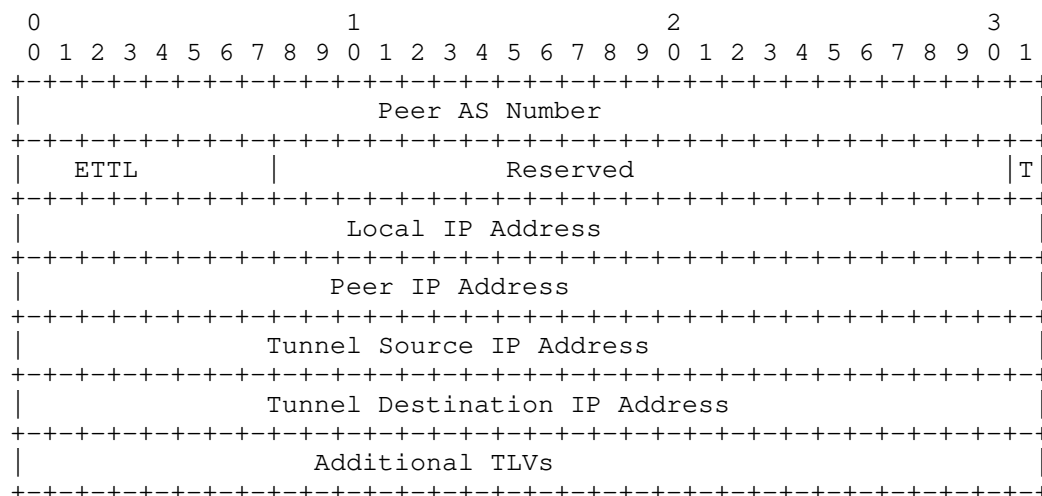


Figure 6: BGP Peer Info Object Body Format for IPv4

[illegible]

ETTL: 1 Byte, to indicate the multihop count for EBGp session. It should be 0 and ignored when Local AS and Peer AS is same.

Reserved: is set to zero while sending, ignored on receipt.

T bit: Indicates whether the traffic that associated with the prefixes advertised via this BGP session is transported via IPinIP tunnel (when T bit is set) or not (when T bit is clear).

Local IP Address(4/16 Bytes): IP address of the local router, used to peer with other end router. When Object-Type is 1, length is 4 bytes; when Object-Type is 2, length is 16 bytes.

Peer IP Address(4/16 Bytes): IP address of the peer router, used to peer with the local router. When Object-Type is 1, length is 4 bytes; when Object-Type is 2, length is 16 bytes;

Tunnel Source IP Address(4/16 Bytes): IP address of the tunnel source, should be owned by the local router. When Object-Type is 1, length is 4 bytes; when Object-Type is 2, length is 16 bytes.

Tunnel Destination IP Address(4/16 Bytes): IP address of the tunnel destination, should be owned by the peer router. When Object-Type is 1, length is 4 bytes; when Object-Type is 2, length is 16 bytes. Should be different from the Peer IP Address.

Additional TLVs: TLVs that associated with this object, can be used to convey other necessary information for dynamic BGP session establishment. Their definition are out of the current document.

When PCC receives BPI object, with Object-Type=1, it should try to establish BGP session with the peer in AFI/SAFI=1/1; when PCC receives BPI object with Object-Type=2, it should try to establish the BGP session with the peer in AFI/SAFI=2/1. Other BGP capabilities, for example, Graceful Restart (GR) that enhance the BGP performance should also be negotiated and used by default.

7.3. Explicit Peer Route Object

The Explicit Peer Route object is defined to specify the explicit peer route to the corresponding peer address on each device that is on the E2E assurance path. This Object should be sent to all the devices that locates on the E2E assurance path that calculated by PCE.

The path established by this object should have higher priority than other path calculated by dynamic IGP protocol, but should be lower priority than the static route configured by manual or NETCONF or by other means.

Explicit Peer Route Object-Class is TBD15.

Explicit Peer Route Object-Type is 1 for IPv4 and 2 for IPv6

The format of Explicit Peer Route object body for IPv4 (Object-Type=1) is as follows:

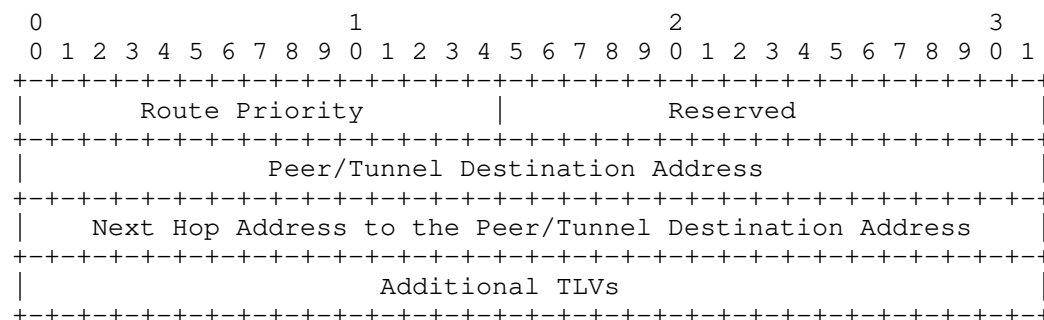


Figure 8: Explicit Peer Route Object Body Format for IPv4

The format of Explicit Peer Route object body for IPv6 (Object-Type=2) is as follows:

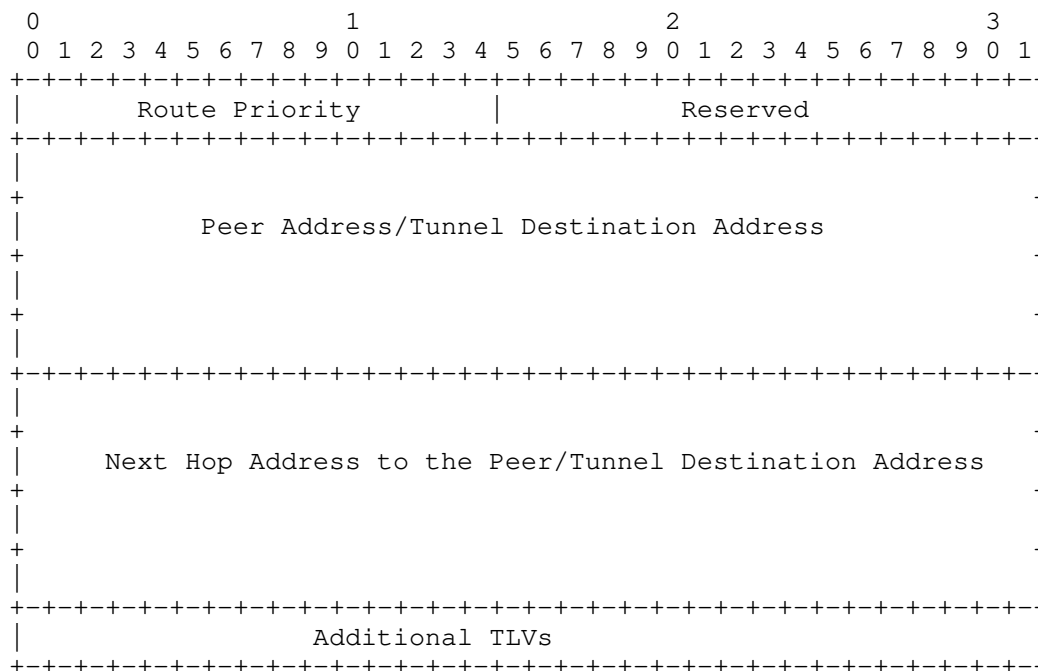


Figure 9: Explicit Peer Route Object Body Format for IPv6

Route Priority: 2 Bytes, The priority of this explicit route. The higher priority should be preferred by the device. This field is used to indicate the backup path at each hop.

Reserved: is set to zero while sending, ignored on receipt.

Peer/Tunnel Destination Address: To indicate the peer address(4/16 Bytes). When T bit is set in the associated BPI object, use the tunnel destination address in BPI object; when T bit is clear, use the peer address in BPI object.

Next Hop Address to the Peer/Tunnel Destination Address: To indicate the next hop address(4/16 Bytes) to the corresponding peer/tunnel destination address.

Additional TLVs: TLVs that associated with this object, can be used to convey other necessary information for explicit peer path establishment. Their definitions are out of the current document.

7.4. Peer Prefix Advertisement Object

The Peer Prefix Advertisement object is defined to specify the IP prefixes that should be advertised to the corresponding peer. This object should only be included and sent to the head/end router of the end2end path.

The prefixes information included in this object MUST only be advertised to the indicated peer, MUST NOT be advertised to other BGP peers.

Peer Prefix Advertisement Object-Class is TBD16

Peer Prefix Advertisement Object-Type is 1 for IPv4 and 2 for IPv6

The format of the Peer Prefix Advertisement object body is as follows:

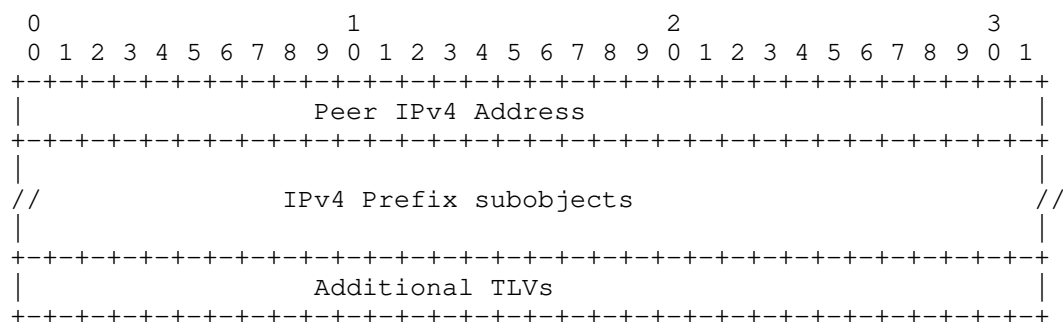


Figure 10: Peer Prefix Advertisement Object Body Format for IPv4

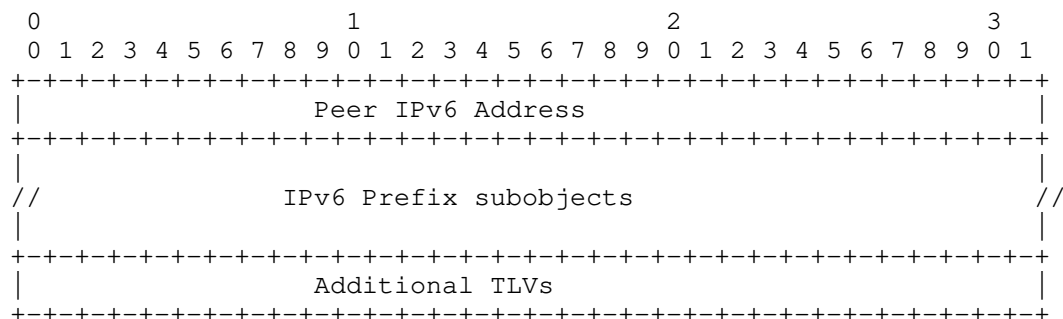


Figure 11: Peer Prefix Advertisement Object Body Format for IPv6

Peer IPv4 Address: 4 Bytes. Identifies the peer IPv4 address that the associated prefixes will be sent to.

IPv4 Prefix subobjects: List of IPv4 Prefix subobjects that defined in [RFC3209], identify the prefixes that will be sent to the peer that identified by Peer IPv4 Address List.

Peer IPv6 Address: 16 Bytes. Identifies the peer IPv6 address that the associated prefixes will be sent to.

IPv6 Prefix subobjects: List of IPv6 Prefix subobjects that defined in [RFC3209], identify the prefixes that will be sent to the peer that identified by Peer IPv6 Address List.

Additional TLVs: TLVs that associated with this object, can be used to convey other necessary information for prefixes advertisement. Their definitions are out of the current document.

8. End to End Path Protection

[RFC8697] defines the path associations procedures between sets of Label Switched Path (LSP). Such procedures can also be used for the E2E path protection. To accomplish this, the PCE should attach the ASSOCIATION object with the EPR object in the PCInitiate message, with the association type set to 1 (Path Protection Association). The Extended Association ID that included within the Extended Association ID TLV, which is included in the ASSOCIATION object, should be set to the Symbolic Path Name of different E2E path. This PCInitiate should be sent to the head-end of the E2E path.

The head-end of the path can use the existing path detection mechanism(for example, Bidirectional Forwarding Detection [RFC5880]), to monitor the status of the active path. Once it detects the failure, it can switch the backup protection path immediately.

9. Re-Delegation and Clean up

In case of a PCE failure, a new PCE can gain control over the central controller instructions. As per the PCEP procedures in [RFC8281], the State Timeout Interval timer is used to ensure that a PCE failure does not result in automatic and immediate disruption for the services. Similarly, as per [RFC9050], the central controller instructions are not removed immediately upon PCE failure. Instead, they could be re-delegated to the new PCE before the expiration of this timer, or be cleaned up on the expiration of this timer. The allows for network clean up without manual intervention. The PCC MUST support the removal of CCI as one of the behaviors applied on expiration of the State Timeout Interval timer.

10. BGP Considerations

This draft defines the procedures and objects to create the BGP sessions and advertises the associated prefixes dynamically. Only the key information, for example peer IP addresses, peer AS number are exchanged via the PCEP protocol. Other parameters that are needed for the BGP session setup should be derived from their default values, as described in Section 7.2. Upon receives such key information, the BGP module on the PCC should try to accomplish the task that appointed by the PCEP protocol and report the status to the PCEP modules.

There is no influence to current implementation of BGP Finite State Machine(FSM). The PCEP cares only the success and failure status of BGP session, and act upon such information accordingly.

The error handling procedures related to incorrect BGP parameters are specified in Section 6.1, Section 6.2, and Section 6.3. The handling of the dynamic BGP sessions and associated prefixes on PCE failure is described in Section 9.

11. New Error-Types and Error-Values Defined

A PCEP-ERROR object is used to report a PCEP error and is characterized by an Error-Type that specifies that type of error and an Error-value that provides additional information about the error. An additional Error-Type and several Error-values are defined to represent some the errors related to the newly defined objects, which are related to Native IP TE procedures.

Error-Type	Meaning	Error-value
TBD6	Native IP TE failure	
		0: Unassigned
		TBD7: Peer AS not match
		TBD8:Peer IP can't be reached
		TBD9:Local IP is in use
		TBD10:Remote IP is in use
		TBD11:Exist BGP session broken
		TBD12:Explicit Peer Route Error
		TBD17:EPR/BPI Peer Info mismatch
		TBD18:BPI/PPA Address Family mismatch
		TBD19:PPA/BPI Peer Info mismatch

Figure 12: Newly defined Error-Type and Error-Value

12. Deployment Considerations

The information transferred in this draft is mainly used for the light weight BGP session setup, explicit route deployment and the prefix distribution. The planning, allocation and distribution of the peer addresses within IGP should be accomplished in advanced and they are out of the scope of this draft.

[RFC8232] describes the state synchronization procedure between stateful PCE and PCC. The communication of PCE and PCC described in this draft should also follow this procedures, treat the three newly defined objects that associated with the same symbolic path name as the attribute of the same path in the LSP-DB.

When PCE detects one or some of the PCCs are out of control, it should recompute and redeploy the traffic engineering path for native IP on the active PCCs. When PCC detects that it is out of control of the PCE, it should clear the information that initiated by the PCE. The PCE should assure the avoidance of possible transient loop in such node failure when it deploy the explicit peer route on the PCCs.

If the established BGP session is broken after some time, the PCC should also report such error via PCErr message with Err-type=TBD6 and error value(Error-value=TBD11, Existing BGP session is broken). Upon receiving such PCErr message, the PCE should clear the prefixes advertisement on the previous BGP session, clear the explicit peer route to the previous peer address; select other Local_IP/Peer_IP pair to establish the new BGP session, deploy the explicit peer route to the new peer address, and advertises the prefixes on the new BGP session.

13. 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".

13.1. Proof of Concept based on ODL

.At the time of posting the -18 version of this document, there are no known implementations of this mechanism. A proof of concept for the overall design has been verified using another SBI protocol on the Open DayLight (ODL) controller.

14. Security Considerations

The setup of BGP sessions, prefix advertisement, and explicit peer route establishment are all controlled by the PCE. See [RFC4271] and [RFC4272] for BGP security considerations. Security consideration part in [RFC5440] and [RFC8231] should be considered. To prevent a bogus PCE sending harmful messages to the network nodes, the network devices should authenticate the validity of the PCE and ensure a secure communication channel between them. Mechanisms described in [RFC8253] should be used.

15. IANA Considerations

15.1. 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	Native IP TE Path	This document

15.2. 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(N)	NATIVE-IP-TE-CAPABILITY	This document

15.3. PCEP Object Types

IANA is requested to allocate new registry for the PCEP Object Type:

Object-Class Value	Name	Reference
44	CCI Object Object-Type TBD13: Native IP	This document
TBD14	BGP Peer Info Object-Type 1: IPv4 address 2: IPv6 address	This document
TBD15	Explicit Peer Route Object-Type 1: IPv4 address 2: IPv6 address	This document
TBD16	Peer Prefix Advertisement Object-Type 1: IPv4 address 2: IPv6 address	This document

15.4. 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:Native IP object missing This document
10	Reception of an invalid object	TBD3:PCECC NATIVE-IP-TE-CAPABILITY bit is not set This document
19	Invalid Operation	TBD5:Only one of the BPI,EPR or PPA object can be included in this message This document
TBD6	Native IP TE failure	This document TBD7:Peer AS not match TBD8:Peer IP can't be reached TBD9:Local IP is in use TBD10:Remote IP is in use TBD11:Exist BGP session broken TBD12:Explicit Peer Route Error TBD17:EPR/BPI Peer Info mismatch TBD18:BPI/PPA Address Family mismatch TBD19:PPA/BPI Peer Info mismatch

16. Contributor

Dhruv Dhody has contributed the contents of this draft.

17. Acknowledgement

Thanks Mike Koldychev, Susan Hares, Siva Sivabalan, Adam Simpson for his valuable suggestions and comments.

18. 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>>.
- [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>>.
- [RFC4271] Rekhter, Y., Ed., Li, T., Ed., and S. Hares, Ed., "A Border Gateway Protocol 4 (BGP-4)", RFC 4271, DOI 10.17487/RFC4271, January 2006, <<https://www.rfc-editor.org/info/rfc4271>>.
- [RFC4272] Murphy, S., "BGP Security Vulnerabilities Analysis", RFC 4272, DOI 10.17487/RFC4272, January 2006, <<https://www.rfc-editor.org/info/rfc4272>>.
- [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>>.
- [RFC5880] Katz, D. and D. Ward, "Bidirectional Forwarding Detection (BFD)", RFC 5880, DOI 10.17487/RFC5880, June 2010, <<https://www.rfc-editor.org/info/rfc5880>>.
- [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>>.

- [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>>.
- [RFC8253] Lopez, D., Gonzalez de Dios, O., Wu, Q., and D. Dhody, "PCEPS: Usage of TLS to Provide a Secure Transport for the Path Computation Element Communication Protocol (PCEP)", RFC 8253, DOI 10.17487/RFC8253, October 2017, <<https://www.rfc-editor.org/info/rfc8253>>.
- [RFC8281] Crabbe, E., Minei, I., Sivabalan, S., and R. Varga, "Path Computation Element Communication Protocol (PCEP) Extensions for PCE-Initiated LSP Setup in a Stateful PCE Model", RFC 8281, DOI 10.17487/RFC8281, December 2017, <<https://www.rfc-editor.org/info/rfc8281>>.
- [RFC8283] Farrel, A., Ed., Zhao, Q., Ed., Li, Z., and C. Zhou, "An Architecture for Use of PCE and the PCE Communication Protocol (PCEP) in a Network with Central Control", RFC 8283, DOI 10.17487/RFC8283, December 2017, <<https://www.rfc-editor.org/info/rfc8283>>.
- [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>>.
- [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>>.
- [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>>.
- [RFC8821] Wang, A., Khasanov, B., Zhao, Q., and H. Chen, "PCE-Based Traffic Engineering (TE) in Native IP Networks", RFC 8821, DOI 10.17487/RFC8821, April 2021, <<https://www.rfc-editor.org/info/rfc8821>>.

[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

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@yahoo.com

Sheng Fang
Huawei Technologies, Co., Ltd
Huawei Bld., No.156 Beiqing Rd.
Beijing
China
Email: fsheng@huawei.com

Ren Tan
Huawei Technologies, Co., Ltd
Huawei Bld., No.156 Beiqing Rd.
Beijing
China
Email: tanren@huawei.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: January 13, 2022

H. Li
A. Wang
China Telecom
H. Chen
Futurewei
R. Chen
ZTE Corporation
July 12, 2021

PCE based BIER Procedures and Protocol Extensions
draft-li-pce-based-bier-01

Abstract

This document describes extensions to Path Computation Element (PCE) communication Protocol (PCEP) for supporting the PCE based Bit Index Explicit Replication (BIER) deployment.

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 January 13, 2022.

Copyright Notice

Copyright (c) 2021 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<https://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of

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

Table of Contents

1. Introduction	2
2. Conventions used in this document	3
3. Terminology	3
4. Overview of PCE based BIER solution	4
4.1. Example of PCE based BIER Topology	4
4.2. Basic Procedures	5
5. Capability Advertisement	5
6. PCEP message	6
6.1. PCRpt message	6
6.2. PCUpd message	7
7. Object formats	8
7.1. Multicast Source Registration Object	8
7.2. Multicast Receiver Information Object	10
7.3. Forwarding Indication Object	11
7.4. Multicast Receiver Status Object	13
8. Procedures	14
8.1. Multicast source registration and revocation	14
8.2. Joining and leaving of multicast receivers	14
8.3. BitString management	15
8.4. Receiver information synchronization	15
9. Deployment Considerations	15
10. Security Considerations	15
11. IANA Considerations	15
11.1. BIER-MULTICAST-CAPABILITY	16
11.2. PCEP-ERROR Object	16
11.3. New Objects	16
12. Contributor	16
13. Acknowledgement	16
14. Normative References	16
Authors' Addresses	18

1. Introduction

[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.

Currently, multicast management information is mainly signaled by PIM [RFC2362] or BGP [RFC6514], which have some limitations in the deployment and process.

[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.

This document specifies PCEP protocol extensions to optimize the implementation of multicast source registration or revocation, receiver automatic discovery, and forwarding control of multicast data by using PCEP messages to transmit multicast management signaling, combining with the forwarding characteristics of BIER.

2. Conventions used in this document

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

3. Terminology

The following terms are used in this document:

- o BFR-id: BFR Identifier. It is a number in the range [1,65535]
- o BGP: Border Gateway Protocol
- o BIER: Bit Index Explicit Replication
- o BIFT: Bit Index Forwarding Table
- o FI: Forwarding indication
- o IGMP: Internet Group Management Protocol
- o IGP: Interior Gateway Protocols
- o MLD: Multicast Listener Discover
- o MRI: Multicast Receiver Information
- o MSR: Multicast Source Registration

- o PCC: Path Computation Client
- o PCE: Path Computation Element
- o PCEP: PCE communication Protocol
- o PIM: Protocol Independent Multicast

4. Overview of PCE based BIER solution

PCE based BIER includes multicast source registration information management, multicast receiver information management and multicast data forwarding control.

Multicast source registration information includes registration and processing of multicast source information.

Multicast receiver information includes requesting multicast group, multicast source and BitPosition information of receiver-side PCC.

Multicast data forwarding control includes BitString processing and data forwarding.

PCRpt message and PCUpd message, described in [RFC8231], are used in the PCE based BIER processing.

This document specifies PCEP protocol extensions for multicast group management, including Multicast Source Registration (MSR) object, Multicast Receiver Information (MRI) object, Forwarding Indication (FI) object and Multicast Receiver Status (MRS) object.

4.1. Example of PCE based BIER Topology

An example of PCE based BIER topology for a BIER domain with a controller as PCE is shown in Figure 1. In this domain, node R1 and R7 are Bit-Forwarding Ingress Router (BFIR) and Bit-Forwarding Egress Router (BFER), respectively.

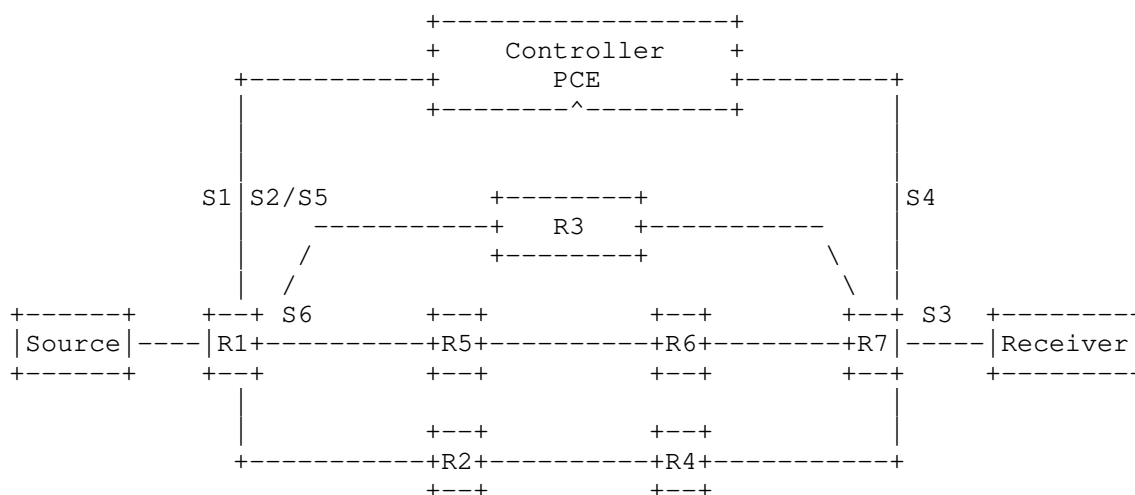


Figure 1: Example of PCE based BIER Topology(controller as PCE)

4.2. Basic Procedures

Step 1(S1): R1 sends multicast source information and authentication information to the controller about multicast information registration via PCRpt message.

Step 2(S2): The controller sends PCUpd message to R1, carrying authentication result.

Step 3(S3): Receivers send IGMP or MLD messages to R7 requesting to join or leave a multicast group.

Step 4(S4): R7 converts the IGMP or MLD messages into PCRpt message and sends it to the controller.

Step 5(S5): If the multicast group and multicast source information requested by the receiver has registered, the controller will send PCUpd message to R1 to start or stop forwarding, carrying BitString.

Step 6(S6): If R1 is ready to start forwarding, it will encapsulate BIER header and forward them based on BIFT and BitString when receiving multicast packets.

5. Capability Advertisement

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:

B (BIER-MULTICAST-CAPABILITY, 1 bit): If set to 1 by a PCEP speaker, it indicates that the PCEP speaker supports the capability of these new flag as specified in this document.

If a PCEP speaker receives PCEP message with the newly defined object, but without the B 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 (BIER-MULTICAST-CAPABILITY bit is not set).
- o Terminate the PCEP session.

6. PCEP message

6.1. PCRpt message

MSR objectSection 7.1 should be included in the PCRpt message when PCC registers multicast source information with PCE.

MRI objectSection 7.2 should be included in the PCRpt message when PCC sends multicast join messages to PCE.

MRS objectSection 7.4 should be included in the PCRpt message when PCC inform PCE of the number of receivers.

The definition of the PCRpt message from [RFC8231] is extended to optionally include MSR object, MRI object and MRS object after the path object. The encoding from [RFC8231] 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>]
                    [<MRS>]
```

Where:

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

6.2. PCUpd message

MSR objectSection 7.1 should be included in the PCUpd message when PCE responds to the registration request.

FI objectSection 7.3 should be included in the PCUpd message when PCE sends the BitString to PCC to indicate the path of multicast data packets forwarding for PCC.

MRS objectSection 7.4 should be included in the PCUpd message when PCE inform PCC of the number of receivers.

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

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

Where:

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

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

Where:

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

7. Object formats

7.1. Multicast Source Registration Object

The MSR object is optional and specifies multicast source information in multicast registration information management. 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:

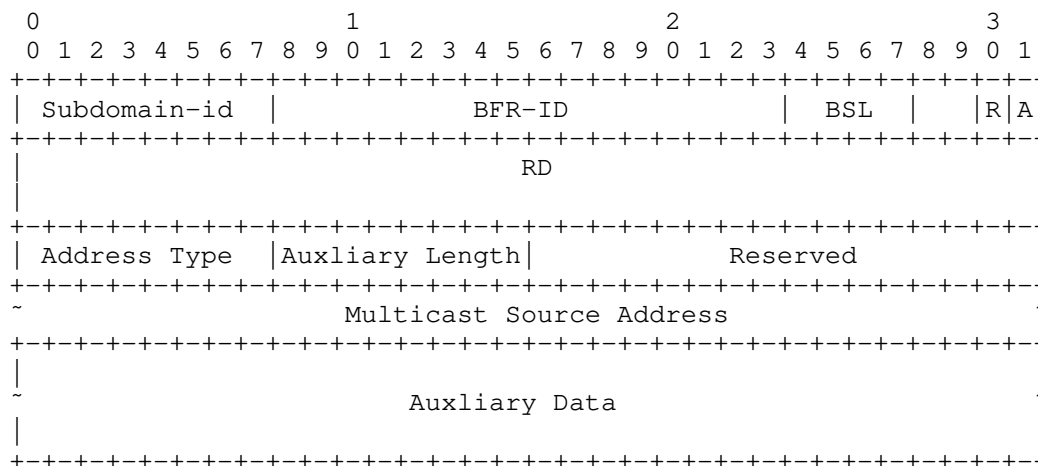


Figure 2: MSR Object Body Format

Subdomain-id(8 bits): Unique value identifying the BIER subdomain.

BFR-ID (16 bits): 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.

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 register.

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 cancellation of registration. R and A cannot both be set to 0 or 1 in PCRpt message.

RD(Route Distinguisher, 8 bytes): indicates the VPN which the receiver used.

Address Type(8 bits): indicates the type of the source address.
Address Type = 1: IPv4 address. Address Type = 2: IPv6 address.

Auxliary Length(8 bits): indicates the length of Auxliary Data.

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

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

Reserved: This field MUST be set to zero on transmission and MUST be ignored on receipt.

7.2. Multicast Receiver Information Object

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

MRI Object-Class is TBD4. MRI Object-Type is 1.

The format of the MRI object body is:

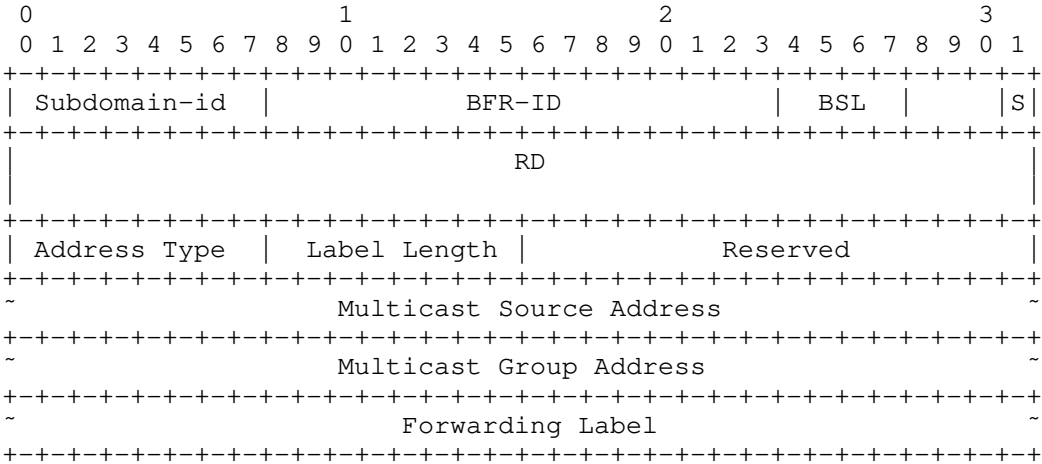


Figure 3: MRI Object Body Format

Subdomain-id(8 bits): Unique value identifying the BIER subdomain.

BFR-ID (16 bits): 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.

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.

RD(Route Distinguisher, 8 bytes): indicates the VPN which the receiver used.

Address Type(8 bits): indicates the type of the source and group addresses. Address Type = 1: IPv4 address. Address Type = 2: IPv6 address.

Label Length(8 bits): indicates the length of Label.

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

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

Forwarding Label(Variable Length): contains MPLS label with 32 bit or IPv6 Segment Identifier with 128 bit.

Reserved: This field MUST be set to zero on transmission and MUST be ignored on receipt.

7.3. Forwarding Indication Object

The FI object is optional and used to indicate to the headend how to forward multicast data packets in the form of BitString. The FI Object should be carried within a PCUpd message sent by PCE to PCC in multicast scenarios.

FI Object-Class is TBD5. FI Object-Type is 1.

The format of the FI object body is:

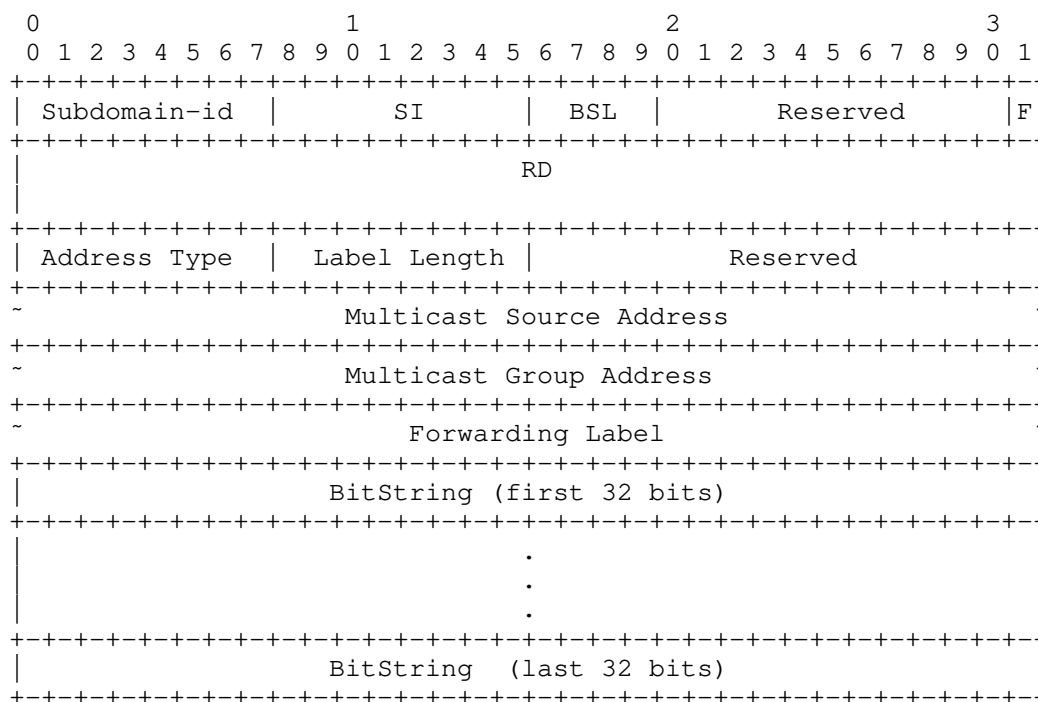


Figure 4: FI Object Body Format

Subdomain-id(8 bits): Unique value identifying the BIER subdomain.

SI (Set Identifier, 8 bits): 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.

F(Forwarding flag, 1 bit): The F flag set to 1 indicates that the router may start forwarding multicast packets. The F flag set to 0 indicates that the router should stop forwarding multicast packets.

RD(Route Distinguisher, 8 bytes): indicates the VPN which the receiver used.

Address Type(8 bits): indicates the type of the source and group addresses. Address Type = 1: IPv4 address. Address Type = 2: IPv6 address.

Label Length(8 bits): indicates the length of Label.

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.

Forwarding Label(Variable Length): contains MPLS label with 32 bit or IPv6 Segment Identifier with 128 bit.

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

Reserved: This field MUST be set to zero on transmission and MUST be ignored on receipt.

7.4. Multicast Receiver Status 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 TBD6. MRS Object-Type is 1.

The format of the MRS object body is:

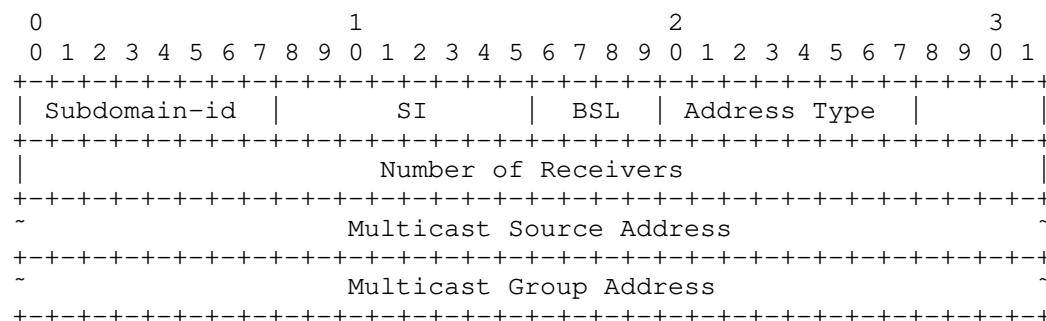


Figure 5: MRS Object Body Format

Subdomain-id(8 bits): Unique value identifying the BIER subdomain.

SI (Set Identifier, 8 bits): 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.

Address Type(8 bits): indicates the type of the source and group addresses. Address Type = 1: IPv4 address. Address Type = 2: IPv6 address.

Number of Receivers(32 bits): indicates the number of receivers for a particular (S,G) tuple.

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.

8. Procedures

8.1. Multicast source registration and revocation

For PCC-Registered multicast source, an ingress node sends a PCRpt message with MSR object to a stateful PCE, where R flag is set and A flag is not set. The registered authentication information can be passed through auxiliary data in MSR object.

Upon receiving the registration via PCRpt message, the stateful PCE MUST match local authentication rules based on the multicast information and auxiliary data in PCRpt message. If authenticated successfully, the PCE stores the multicast registration information into the database. In response, PCE MUST send a PCUpd message with MSR object to ingress node, where R flag is set. A flag is set only if authentication is successful.

For PCC-revoked multicast source registration, an ingress node sends a PCRpt message with MSR object to a stateful PCE, where R flag is not set and A flag is set.

Upon receiving the revocation via PCRpt message, in response, PCE MUST send a PCUpd message with MSR object to ingress node, where neither R nor A is set.

8.2. Joining and leaving of multicast receivers

When an egress node receives an IGMP or MLD message from a multicast receiver to join, the egress node should send a PCRpt message with MRI object to the PCE if no other receiver has sent the same request to it before.

If it is not the first time the PCE has received the same PCRpt message for join from the same egress node, this message should be ignored.

When an egress node receives an IGMP or MLD message from a multicast receiver to leave, the egress node should send a PCRpt message with MRI object and MRS object to the PCE if there are no other members in the requested multicast group. In MRS object, the number of receivers is zero.

8.3. BitString management

Upon receiving the join or leave request via PCRpt message, PCE needs to combine the BFR-id and SI of the egress node carried in PCRpt message with the BFR-id and SI of the ingress node and existed BitStrings in the database to create or update BitString. If there are members in the multicast group, the PCE should send a PCUpd message with FI object carrying the latest BitString to the ingress node, where F flag is set.

When receiving multicast packets, the ingress node encapsulates BIER header and forwards them based on BIFT and BitString. Encapsulation of Forwarding Label is not in the scope of this document.

If there is no member in the multicast group, the PCE should send a PCUpd message with FI object to the ingress node, where F flag is not set.

8.4. Receiver information synchronization

Upon receiving multicast packets from a particular multicast group, egress node will synchronize the number of receivers in this multicast group with the PCE via PCRpt message with MRS object periodically.

After sending a PCUpd message with FI object to an ingress node for a particular multicast group, the PCE will synchronize the total number of receivers in this multicast group with the ingress node via PCUpd message with MRS object periodically.

If there is no member in the multicast group, the synchronization of receiver number information ends.

9. Deployment Considerations

10. Security Considerations

11. IANA Considerations

11.1. BIER-MULTICAST-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	BIER-MULTICAST-CAPABILITY	This document

11.2. PCEP-ERROR Object

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

Error-Type	Description	Reference
10	Error-value = TBD2 B bit is not set	This document

11.3. New Objects

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

Object-Class Value	Description	Reference
TBD3	Multicast Receiver Information	This document
TBD4	Multicast Receiver Information	This document
TBD5	Forwarding Indication	This document
TBD6	Multicast Receiver Status	This document

12. Contributor

13. Acknowledgement

14. 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>>.

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

Authors' Addresses

Huanan Li
China Telecom
Beiqijia Town, Changping District
Beijing, Beijing 102209
China

Email: lihn6@foxmail.com

Aijun Wang
China Telecom
Beiqijia Town, Changping District
Beijing, Beijing 102209
China

Email: wangaj3@chinatelecom.cn

Huaimo Chen
Futurewei
Boston
USA

Email: Huaimo.chen@futurewei.com

Ran Chen
ZTE Corporation
50 Software Avenue, Yuhua District
Nanjing, Jiangsu 210012
China

Email: chen.ran@zte.com.cn

PCE Working Group
Internet-Draft
Intended status: Standards Track
Expires: April 22, 2022

H. Li
A. Wang
China Telecom
H. Chen
Futurewei
R. Chen
ZTE Corporation
October 19, 2021

PCE based BIER Procedures and Protocol Extensions
draft-li-pce-based-bier-02

Abstract

This document describes extensions to Path Computation Element (PCE) communication Protocol (PCEP) for supporting the PCE based Bit Index Explicit Replication (BIER) deployment.

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 April 22, 2022.

Copyright Notice

Copyright (c) 2021 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<https://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of

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

Table of Contents

1. Introduction	2
2. Conventions used in this document	3
3. Terminology	3
4. Overview of PCE based BIER solution	4
4.1. Example of PCE based BIER Topology	4
4.2. Basic Procedures	5
5. Capability Advertisement	5
6. PCEP message	6
6.1. PCRpt message	6
6.2. PCUpd message	7
7. Object formats	8
7.1. Multicast Source Registration Object	8
7.1.1. Multicast Source Address TLV	9
7.1.2. BIER Information TLV	10
7.1.3. VPN Information TLV	10
7.2. Multicast Receiver Information Object	11
7.2.1. Multicast Group Address TLV	12
7.3. Forwarding Indication Object	12
7.4. Multicast Receiver Status Object	13
8. Procedures	14
8.1. Multicast source registration and revocation	14
8.2. Joining and leaving of multicast receivers	15
8.3. BitString management	15
8.4. Receiver information synchronization	15
9. Deployment Considerations	16
10. Security Considerations	16
11. IANA Considerations	16
11.1. BIER-MULTICAST-CAPABILITY	16
11.2. PCEP-ERROR Object	16
11.3. New Objects	16
11.4. New TLVs	16
12. Contributor	17
13. Acknowledgement	17
14. Normative References	17
Authors' Addresses	18

1. Introduction

[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.

Currently, multicast management information is mainly signaled by PIM [RFC2362] or BGP [RFC6514], which have some limitations in the deployment and process.

[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.

This document specifies PCEP protocol extensions to optimize the implementation of multicast source registration or revocation, receiver automatic discovery, and forwarding control of multicast data by using PCEP messages to transmit multicast management signaling, combining with the forwarding characteristics of BIER.

2. Conventions used in this document

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

3. Terminology

The following terms are used in this document:

- o BFR-id: BFR Identifier. It is a number in the range [1,65535]
- o BGP: Border Gateway Protocol
- o BIER: Bit Index Explicit Replication
- o BIFT: Bit Index Forwarding Table
- o FI: Forwarding indication
- o IGMP: Internet Group Management Protocol
- o IGP: Interior Gateway Protocols
- o MLD: Multicast Listener Discover

- o MRI: Multicast Receiver Information
- o MSR: Multicast Source Registration
- o PCC: Path Computation Client
- o PCE: Path Computation Element
- o PCEP: PCE communication Protocol
- o PIM: Protocol Independent Multicast

4. Overview of PCE based BIER solution

PCE based BIER includes multicast source registration information management, multicast receiver information management and multicast data forwarding control.

Multicast source registration information includes registration and processing of multicast source information.

Multicast receiver information includes requesting multicast group, multicast source and BitPosition information of receiver-side PCC.

Multicast data forwarding control includes BitString processing and data forwarding.

PCRpt message and PCUpd message, described in [RFC8231], are used in the PCE based BIER processing.

This document specifies PCEP protocol extensions for multicast group management, including Multicast Source Registration (MSR) object, Multicast Receiver Information (MRI) object, Forwarding Indication (FI) object and Multicast Receiver Status (MRS) object.

4.1. Example of PCE based BIER Topology

An example of PCE based BIER topology for a BIER domain with a controller as PCE is shown in Figure 1. In this domain, node R1 and R7 are Bit-Forwarding Ingress Router (BFIR) and Bit-Forwarding Egress Router (BFER), respectively.

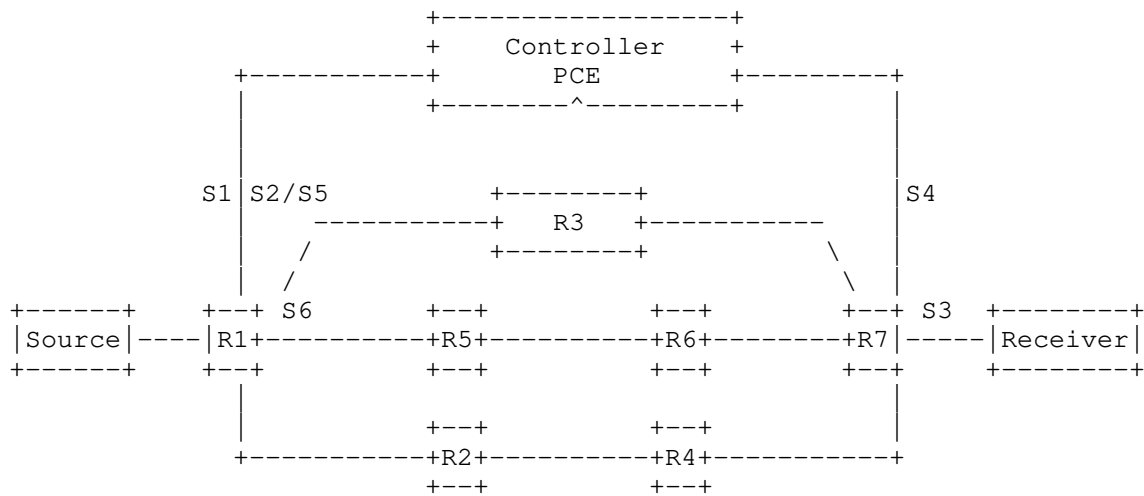


Figure 1: Example of PCE based BIER Topology(controller as PCE)

4.2. Basic Procedures

Step 1(S1): R1 sends multicast source information and authentication information to the controller about multicast information registration via PCRpt message.

Step 2(S2): The controller sends PCUpd message to R1, carrying authentication result.

Step 3(S3): Receivers send IGMP or MLD messages to R7 requesting to join or leave a multicast group.

Step 4(S4): R7 converts the IGMP or MLD messages into PCRpt message and sends it to the controller.

Step 5(S5): If the multicast group and multicast source information requested by the receiver has registered, the controller will send PCUpd message to R1 to start or stop forwarding, carrying BitString.

Step 6(S6): If R1 is ready to start forwarding, it will encapsulate BIER header and forward them based on BIFT and BitString when receiving multicast packets.

5. Capability Advertisement

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:

B (BIER-MULTICAST-CAPABILITY, 1 bit): If set to 1 by a PCEP speaker, it indicates that the PCEP speaker supports the capability of these new flag as specified in this document.

If a PCEP speaker receives PCEP message with the newly defined object, but without the B 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 (BIER-MULTICAST-CAPABILITY bit is not set).
- o Terminate the PCEP session.

6. PCEP message

6.1. PCRpt message

MSR objectSection 7.1 should be included in the PCRpt message when PCC registers multicast source information with PCE.

MRI objectSection 7.2 should be included in the PCRpt message when PCC sends multicast join messages to PCE.

MRS objectSection 7.4 should be included in the PCRpt message when PCC inform PCE of the number of receivers.

The definition of the PCRpt message from [RFC8231] is extended to optionally include MSR object, MRI object and MRS object after the path object. The encoding from [RFC8231] 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>]
                   [<MRS>]
```

Where:

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

6.2. PCUpd message

MSR objectSection 7.1 should be included in the PCUpd message when PCE responds to the registration request.

FI objectSection 7.3 should be included in the PCUpd message when PCE sends the BitString to PCC to indicate the path of multicast data packets forwarding for PCC.

MRS objectSection 7.4 should be included in the PCUpd message when PCE inform PCC of the number of receivers.

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

```

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

```

Where:

```

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

```

```

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

```

Where:

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

7. Object formats

7.1. Multicast Source Registration Object

The MSR object is optional and specifies multicast source information in multicast registration information management. 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:

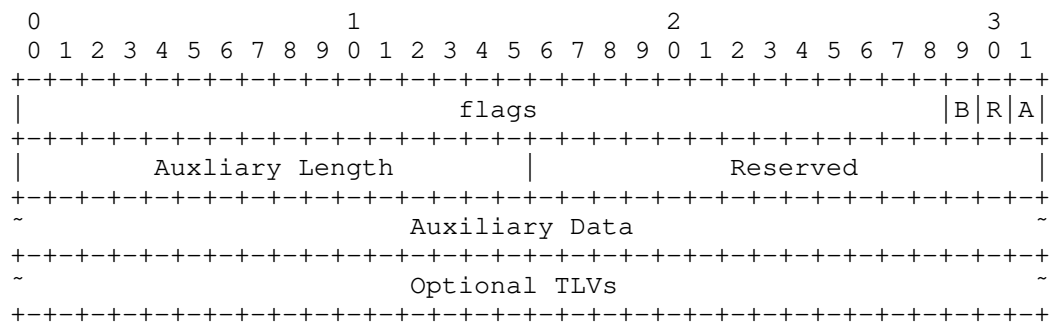


Figure 2: MSR Object Body Format

B(BIER multicast flag, 1 bit): The R flag set to 1 indicates that multicast protocol is BIER. The R flag set to 0 indicates that multicast protocol is not BIER.

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 register.

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 cancellation of registration. R and A cannot both be set to 0 or 1 in PCRpt message.

Auxiliary Length(8 bits): indicates the length of Auxiliary Data.

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

MSR object could include three types of TLVs, namely Multicast Source Address TLV, BIER Information TLV, VPN Information TLV, as defined follows:

7.1.1. Multicast Source Address TLV

The format of the Multicast Source Address TLV is:

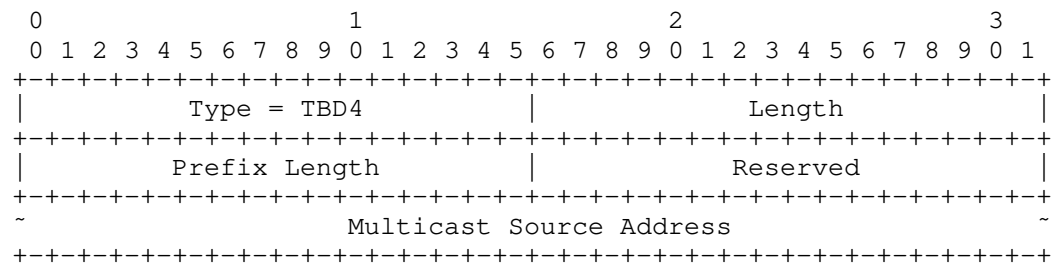


Figure 3: Multicast Source Address TLV Format

Type(16 bits): TBD4 is to be assigned by IANA.

Length: Variable.

Prefix Length(16 bits): indicates the length of multicast source address.

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

7.1.2. BIER Information TLV

BIER Information TLV is used to report router location information in the BIER domain. When the multicast flag in MSR, MRI, FI objects is set, BIER Information TLV should be included. The format of the BIER Information TLV is:

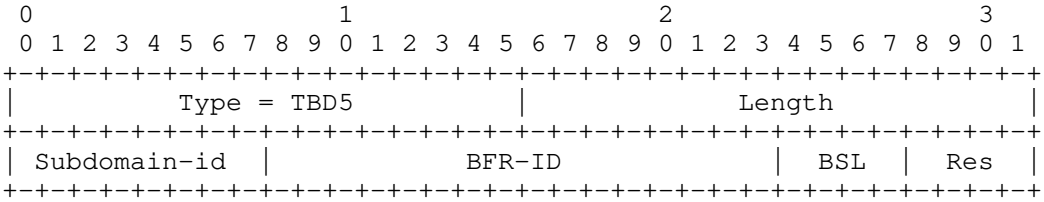


Figure 4: BIER Information TLV Format

Type(16 bits): TBD5 is to be assigned by IANA.

Length: Variable.

Subdomain-id(8 bits): Unique value identifying the BIER subdomain.

BFR-ID (16 bits): 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.

7.1.3. VPN Information TLV

VPN Information TLV is used to report VPN information about multicast sources and receivers. When the multicast flag in MSR, MRI, FI objects is set, VPN Information TLV should be included. The format of the VPN Information TLV is:

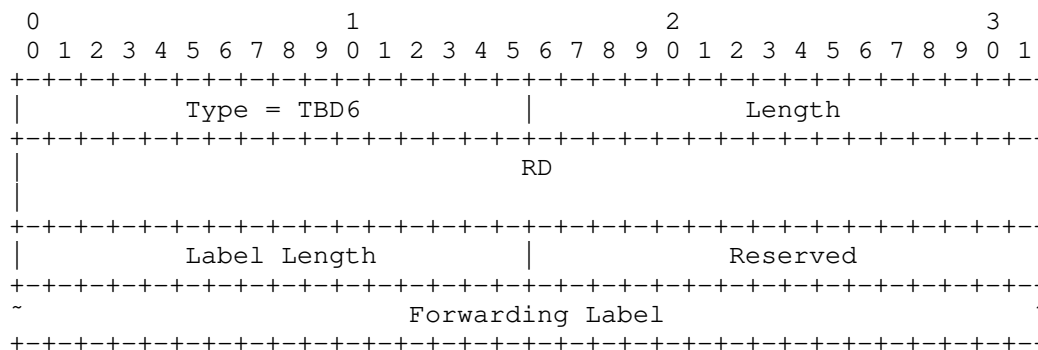


Figure 5: VPN Information TLV Format

Type(16 bits): TBD6 is to be assigned by IANA.

Length: Variable.

RD(Route Distinguisher, 8 bytes): indicates the VPN which the receiver used.

Label Length(16 bits): indicates the length of forwarding label Data, the length should be 0 ,32 bits or 128 bits.

Forwarding Label(Variable Length): contains MPLS label with 32 bit or IPv6 Segment Identifier with 128 bits.

7.2. Multicast Receiver Information Object

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

MRI Object-Class is TBD7. MRI Object-Type is 1.

The format of the MRI object body is:

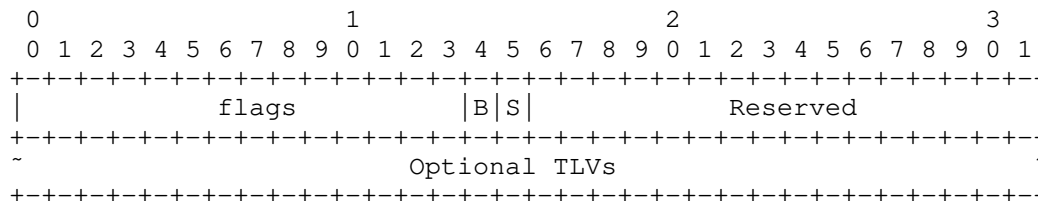


Figure 6: MRI Object Body Format

B(BIER multicast flag, 1 bit): The R flag set to 1 indicates that multicast protocol is BIER. The R flag set to 0 indicates that multicast protocol is not BIER.

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 could include four types of TLVs, namely Multicast Source Address TLV Section 7.1.1, BIER INFO TLV Section 7.1.2, VPN Information TLV Section 7.1.3 and Multicast Group Address TLV. Multicast Group Address TLV is defined as follows:

7.2.1. Multicast Group Address TLV

The format of the Multicast Group Address TLV is:

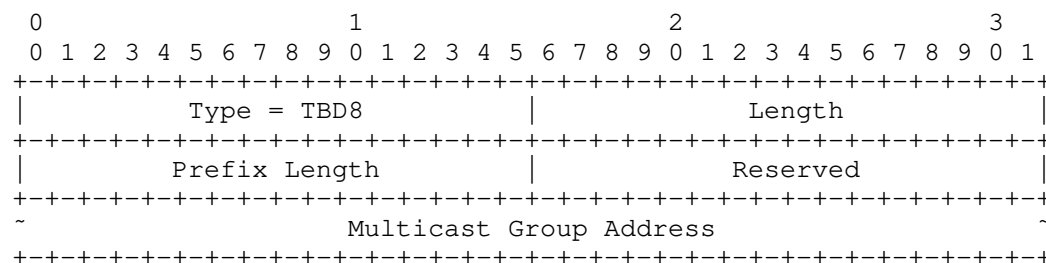


Figure 7: Multicast Group Address TLV Format

Type(16 bits): TBD8 is to be assigned by IANA.

Length: Variable.

Prefix Length(16 bits): indicates the length of multicast group address.

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

7.3. Forwarding Indication Object

The FI object is optional and used to indicate to the headend how to forward multicast data packets in the form of BitString. The FI object should be carried within a PCUpd message sent by PCE to PCC in multicast scenarios.

FI Object-Class is TBD9. FI Object-Type is 1.

The format of the FI object body is:

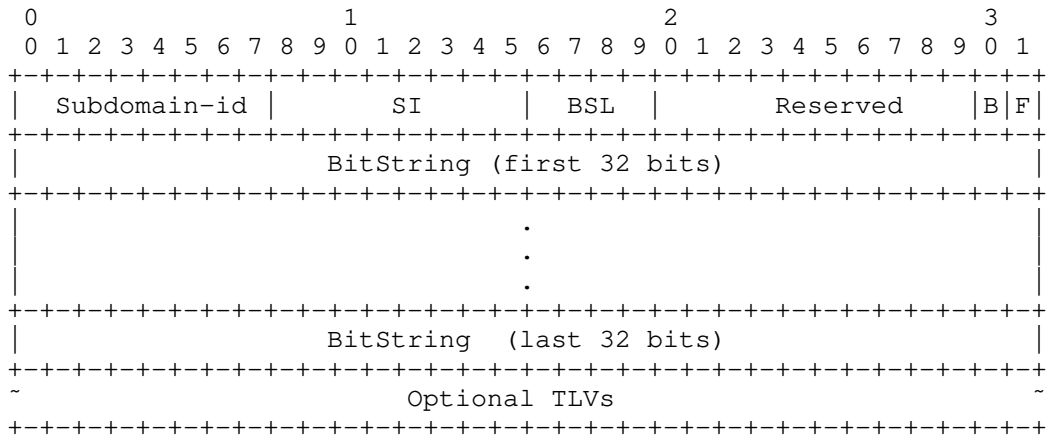


Figure 8: FI Object Body Format

Subdomain-id(8 bits): Unique value identifying the BIER subdomain.

SI (Set Identifier, 8 bits): 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.

B(BIER multicast flag, 1 bit): The R flag set to 1 indicates that multicast protocol is BIER. The R flag set to 0 indicates that multicast protocol is not BIER.

F(Forwarding flag, 1 bit): The F flag set to 1 indicates that the router may start forwarding multicast packets. The F flag set to 0 indicates that the router should stop forwarding multicast packets.

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

FI object should include three types of TLVs, namely Multicast Source Address TLVSection 7.1.1, VPN Information TLVSection 7.1.3 and Multicast Group Address TLVSection 7.2.1.

7.4. Multicast Receiver Status 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 TBD10. MRS Object-Type is 1.

The format of the MRS object body is:

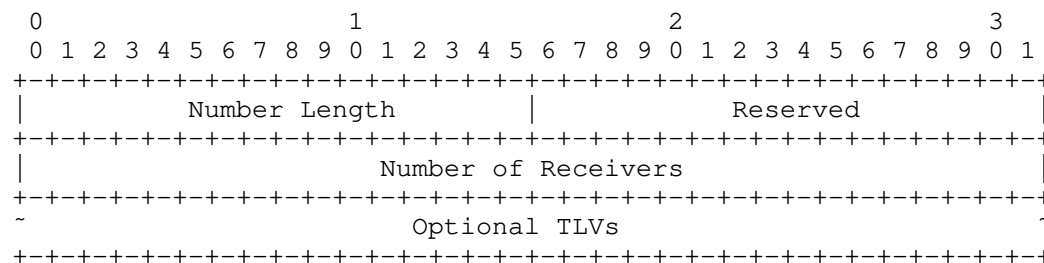


Figure 9: MRS Object Body Format

Number Length(16 bits): indicates the length of receiver number.

Number of Receivers(32 bits): indicates the number of receivers for a particular (S,G) tuple.

MRS object should include two types of TLVs, namely Multicast Source Address TLV Section 7.1.1 and Multicast Group Address TLV Section 7.2.1.

8. Procedures

8.1. Multicast source registration and revocation

For PCC-Registered multicast source, an ingress node sends a PCRpt message with MSR object to a stateful PCE, where R flag is set and A flag is not set. The registered authentication information can be passed through auxiliary data in MSR object.

Upon receiving the registration via PCRpt message, the stateful PCE MUST match local authentication rules based on the multicast information and auxiliary data in PCRpt message. If authenticated successfully, the PCE stores the multicast registration information into the database. In response, PCE MUST send a PCUpd message with MSR object to ingress node, where R flag is set. A flag is set only if authentication is successful.

For PCC-revoked multicast source registration, an ingress node sends a PCRpt message with MSR object to a stateful PCE, where R flag is not set and A flag is set.

Upon receiving the revocation via PCRpt message, in response, PCE MUST send a PCUpd message with MSR object to ingress node, where neither R nor A is set.

8.2. Joining and leaving of multicast receivers

When an egress node receives an IGMP or MLD message from a multicast receiver to join, the egress node should send a PCRpt message with MRI object to the PCE if no other receiver has sent the same request to it before.

If it is not the first time the PCE has received the same PCRpt message for join from the same egress node, this message should be ignored.

When an egress node receives an IGMP or MLD message from a multicast receiver to leave, the egress node should send a PCRpt message with MRI object and MRS object to the PCE if there are no other members in the requested multicast group. In MRS object, the number of receivers is zero.

8.3. BitString management

Upon receiving the join or leave request via PCRpt message, PCE needs to combine the BFR-id and SI of the egress node carried in PCRpt message with the BFR-id and SI of the ingress node and existed BitStrings in the database to create or update BitString. If there are members in the multicast group, the PCE should send a PCUpd message with FI object carrying the latest BitString to the ingress node, where F flag is set.

When receiving multicast packets, the ingress node encapsulates BIER header and forwards them based on BIFT and BitString. Encapsulation of Forwarding Label is not in the scope of this document.

If there is no member in the multicast group, the PCE should send a PCUpd message with FI object to the ingress node, where F flag is not set.

8.4. Receiver information synchronization

Upon receiving multicast packets from a particular multicast group, egress node will synchronize the number of receivers in this multicast group with the PCE via PCRpt message with MRS object periodically.

After sending a PCUpd message with FI object to an ingress node for a particular multicast group, the PCE will synchronize the total number of receivers in this multicast group with the ingress node via PCUpd message with MRS object periodically.

If there is no member in the multicast group, the synchronization of receiver number information ends.

9. Deployment Considerations

10. Security Considerations

11. IANA Considerations

11.1. BIER-MULTICAST-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	BIER-MULTICAST-CAPABILITY	This document

11.2. PCEP-ERROR Object

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

Error-Type	Description	Reference
10	Error-value = TBD2 B bit is not set	This document

11.3. New Objects

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

Object-Class Value	Description	Reference
TBD3	Multicast Receiver Information	This document
TBD7	Multicast Receiver Information	This document
TBD9	Forwarding Indication	This document
TBD10	Multicast Receiver Status	This document

11.4. New TLVs

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

Type	Description	Reference
TBD4	Multicast Source Address	This document
TBD5	Multicast Group Address	This document
TBD6	BIER Information TLV	This document
TBD8	VPN Information	This document

12. Contributor

13. Acknowledgement

14. 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>>.
- [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>>.
- [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>>.
- [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>>.

- [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>>.

Authors' Addresses

Huanan Li
China Telecom
Beiqijia Town, Changping District
Beijing, Beijing 102209
China

Email: lihn6@foxmail.com

Aijun Wang
China Telecom
Beiqijia Town, Changping District
Beijing, Beijing 102209
China

Email: wangaj3@chinatelecom.cn

Huaimo Chen
Futurewei
Boston
USA

Email: Huaimo.chen@futurewei.com

Ran Chen
ZTE Corporation
50 Software Avenue, Yuhua District
Nanjing, Jiangsu 210012
China

Email: chen.ran@zte.com.cn

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

Q. Xiong
S. Peng
ZTE Corporation
F. Qin
China Mobile
July 7, 2021

PCEP Extension for SR-MPLS Entropy Label Position
draft-peng-pce-entropy-label-position-06

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 January 8, 2022.

Copyright Notice

Copyright (c) 2021 IETF Trust and the persons identified as the document authors. All rights reserved.

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

Table of Contents

1. Introduction	2
2. Conventions used in this document	3
2.1. Terminology	3
2.2. Requirements Language	3
3. Entropy Labels in SR-MPLS Scenario with PCE	3
4. PCEP Extensions	5
4.1. The OPEN Object	5
4.2. The LSP-EXTENDED-FLAG TLV	5
4.3. The PATH-MINIMUM-ERLD TLV	6
4.4. The SR-ERO Object	6
5. Operations	7
6. Security Considerations	7
7. Acknowledgements	7
8. IANA Considerations	7
8.1. New SR PCE Capability Flag Registry	7
8.2. New LSP-EXTENDED-FLAG Flag Registry	8
8.3. The PATH-MINIMUM-ERLD TLV	8
8.4. New SR-ERO Flag Registry	8
9. Normative References	9
Authors' Addresses	10

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 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. But in some cases, 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:

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

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.

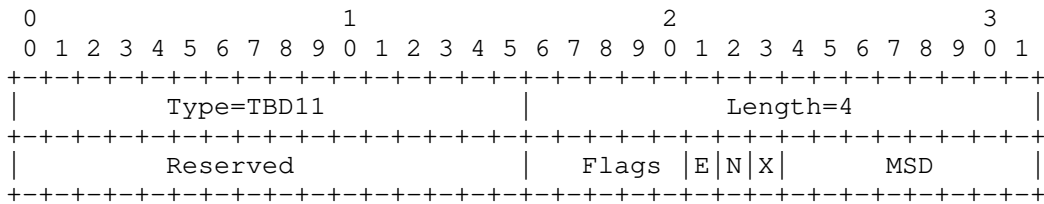


Figure 2: E-flag in SR-PCE-CAPABILITY sub-TLV

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 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:

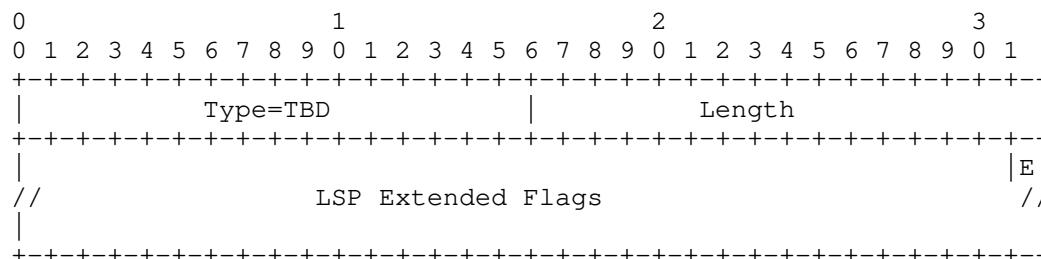


Figure 3: E-flag in LSP-EXTENDED-FLAG TLV

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 PATH-MINIMUM-ERLD TLV

The PATH-MINIMUM-ERLD TLV is an optional TLV for use in the LSP Object for the path minimum ERLD configuration. The type of this TLV is to be allocated by IANA. The format is as shown below.

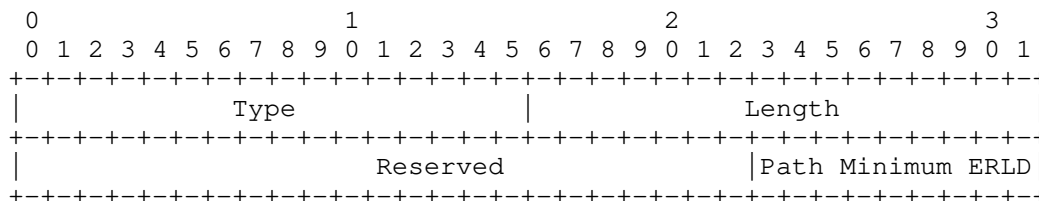


Figure 4: The PATH-MINIMUM-ERLD TLV

Path Minimum ERLD: 8 bits, indicates the minimum ERLD of the nodes along the path.

4.4. 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:

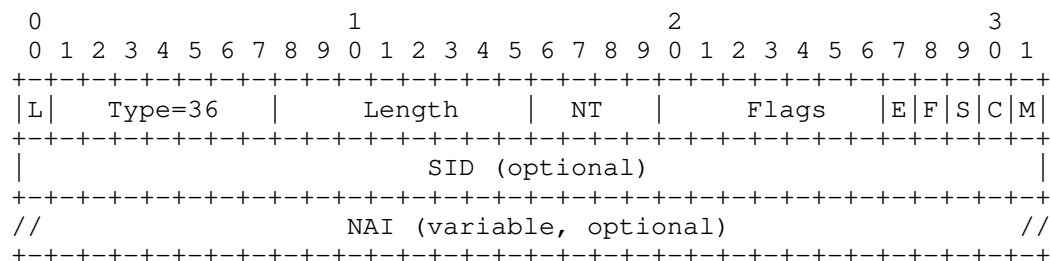


Figure 5: E-flag in SR-ERO subobject

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 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 MUST be inserted into the label stack of the SR-TE 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

TBA

7. Acknowledgements

TBA

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. The PATH-MINIMUM-ERLD TLV

This document requests that a new sub-registry named "PATH-MINIMUM-ERLD TLV" carried in LSP object.

Value	Name	Reference
TBD	PATH-MINIMUM-ERLD TLV	[this document]

Table 3

8.4. 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 4

9. Normative References

- [I-D.ietf-pce-lsp-extended-flags]
Xiong, Q., "LSP Object Flag Extension of Stateful PCE",
draft-ietf-pce-lsp-extended-flags-00 (work in progress),
March 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>>.
- [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>>.

- [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: 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.

Copyright Notice

Copyright (c) 2022 IETF Trust and the persons identified as the document authors. All rights reserved.

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

Table of Contents

1. Introduction	2
2. Conventions used in this document	3
2.1. Terminology	3
2.2. Requirements Language	3
3. Entropy Labels in SR-MPLS Scenario with PCE	3
4. PCEP Extensions	5
4.1. The OPEN Object	5
4.2. The LSP-EXTENDED-FLAG TLV	5
4.3. The SR-ERO Object	6
5. Operations	7
6. Security Considerations	7
7. Acknowledgements	7
8. IANA Considerations	7
8.1. New SR PCE Capability Flag Registry	7
8.2. New LSP-EXTENDED-FLAG Flag Registry	7
8.3. New SR-ERO Flag Registry	8
9. Normative References	8
Authors' Addresses	10

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 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 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 within 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.

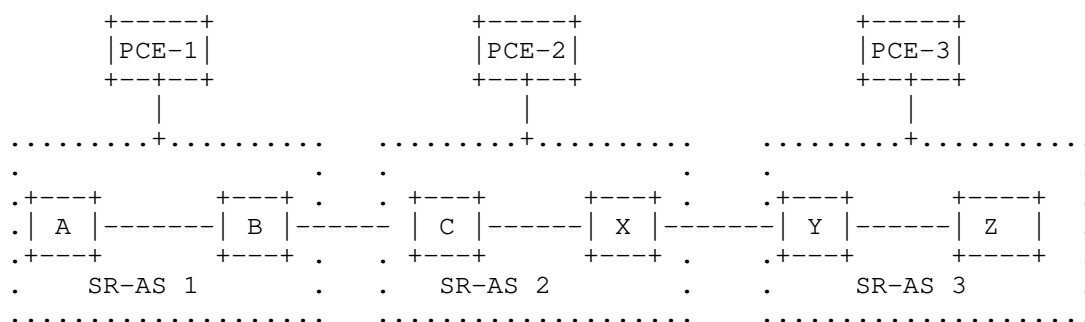


Figure 1: Figure 1: Entropy Labels in SR-MPLS Inter-Domain Scenario

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.

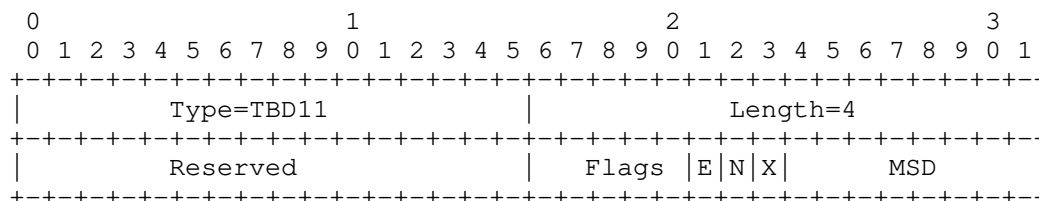


Figure 2: Figure 2: E-flag in SR-PCE-CAPABILITY sub-TLV

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 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:

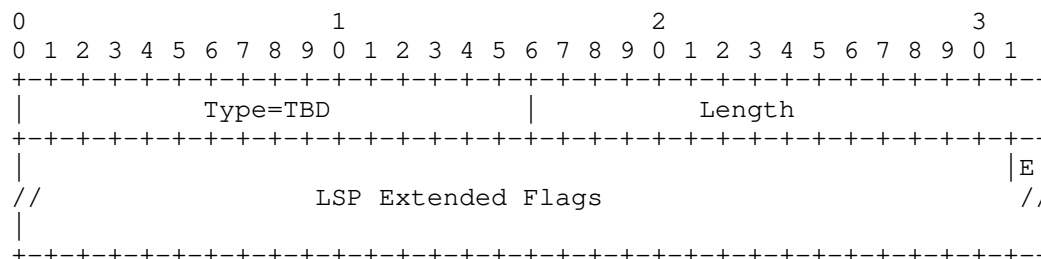


Figure 3: Figure 3: E-flag in LSP-EXTENDED-FLAG TLV

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:

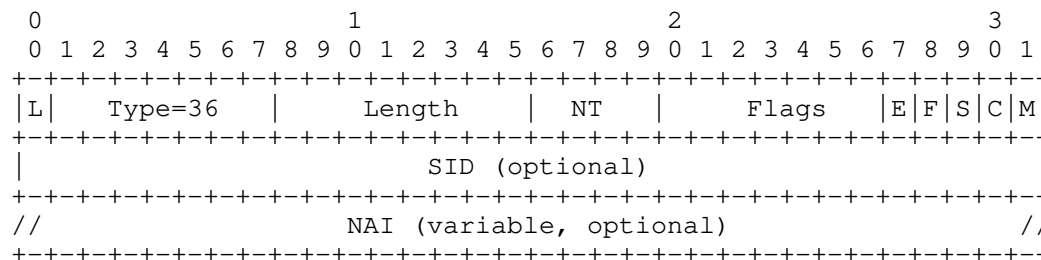


Figure 4: Figure 4: E-flag in SR-ERO subobject

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-srmppls-elp]
Liu, Y. and S. Peng, "BGP Extension for SR-MPLS Entropy Label Position", Work in Progress, Internet-Draft, draft-zhou-idr-bgp-srmppls-elp-04, 1 March 2022, <<https://www.ietf.org/archive/id/draft-zhou-idr-bgp-srmppls-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: January 12, 2022

S. Peng
Q. Xiong
ZTE Corporation
F. Qin
China Mobile
M. Koldychev
Cisco Systems
S. Sivabalan
Ciena Corporation
July 11, 2021

PCE TE Constraints
draft-peng-pce-te-constraints-06

Abstract

This document proposes a set of extensions for PCEP to support the TE constraints during path computation, e.g, IGP instance, virtual network, Slice-id, specific application, color template and FA-id etc.

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 January 12, 2022.

Copyright Notice

Copyright (c) 2021 IETF Trust and the persons identified as the document authors. All rights reserved.

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

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

Table of Contents

1. Introduction	2
2. Conventions used in this document	3
2.1. Terminology	3
2.2. Requirements Language	3
3. PCEP Extensions for TE Constraints	3
3.1. Source Protocol TLV	3
3.2. Multi-topology TLV	4
3.3. Slice-id TLV	5
3.4. Application Specific TLV	6
3.5. Color TLV	7
3.6. FA-id TLV	9
4. Security Considerations	10
5. Acknowledgements	10
6. IANA Considerations	10
7. Normative References	11
Authors' Addresses	13

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. As depicted in [RFC4655], a PCE MUST be able to compute the path of a TE LSP by operating on the TED and considering bandwidth and other constraints applicable to the TE LSP service request. The constraint parameters are provided such as metric, bandwidth, delay, affinity, etc. However these parameters can't meet the network slicing requirements.

A PCE always perform path computation based on the network topology information collected through BGP-LS [RFC7752]. BGP-LS can get multiple link-state data from multiple IGP instance, or multiple virtual topologies from a single IGP instance. It is necessary to restrict the PCE to a small topology scope during path computation for some special purpose. BGP-LS can also get application specific TE attributes for a link, it is also necessary to restrict PCE to use

TE attributes of specific application. The PCE MUST take the identifier of slicing into consideration during path computation.

This document proposes a set of extensions for PCEP to support the TE constraints during path computation, e.g, IGP instance, virtual network, Slice-id, specific application, color template and FA-id etc.

2. Conventions used in this document

2.1. Terminology

The terminology is defined as [RFC5440] and [RFC7752].

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. PCEP Extensions for TE Constraints

As defined in [RFC5440], the LSPA object is used to specify the LSP attributes to be taken into account by the PCE during path computation such as TE constraints. This document proposes several new TLVs for the LSPA object to carry TE constraints in Network Slicing.

3.1. Source Protocol TLV

The Source Protocol TLV is optional and is defined to carry the source protocol constraint.

In a PCReq/PCRpT message, a PCC MAY insert one or more Source Protocol TLVs to indicate the source protocol that MUST be considered by the PCE. If more than one Source Protocol TLVs are carried, the PCE may perform path computation based on the sub-topology identified by the one of the source protocols. The absence of the Source Protocol TLV MUST be interpreted by the PCE as a path computation request for which no constraints need be applied to any of the source protocols.

In a PCRep/PCInit/PCUpd message, the Source Protocol TLV MAY be carried so as to provide the source protocol information for the computed path.

The format of the Source Protocol TLV is shown as Figure 1:

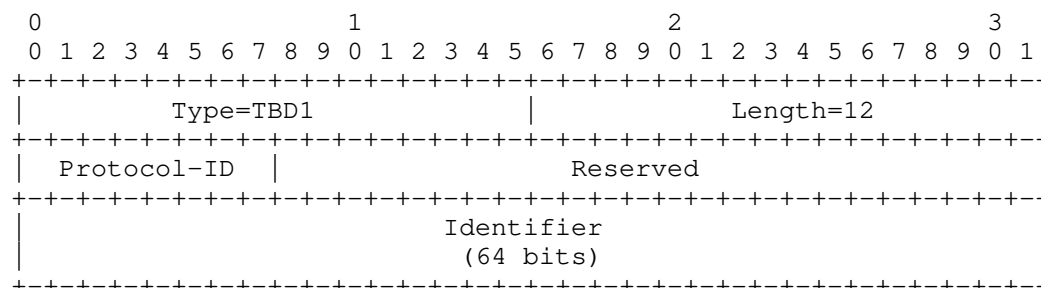


Figure 1: Source Protocol TLV

The code point for the TLV type is TBD1. The TLV length is 12 octets.

Protocol-ID (8 bits): defined in [RFC7752] section 3.2.

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

Identifier (64 bits): defined in [RFC7752] section 3.2.

3.2. Multi-topology TLV

The Multi-topology TLV is optional and is defined to carry the multi-topology protocol constraint.

In a PCReq message, a PCC MAY insert one Multi-topology TLV to indicate the sub-topology of an IGP instance that MUST be considered by the PCE. The PCE will perform path computation based on the sub-topology identified by the specific Multi-Topology ID within a source protocol. The absence of the Multi-topology TLV MUST be interpreted by the PCE as a path computation request for which no constraints need be applied to any of the multi-topologies.

In a PCRep/PCInit/PCUpd message, the Multi-topology TLV MAY be carried so as to provide the Multi-topology information for the computed path.

The Multi-topology TLV MUST be carried after a Source Protocol TLV, if not it MUST be ignored.

The format of the Multi-topology TLV is shown as Figure 2:

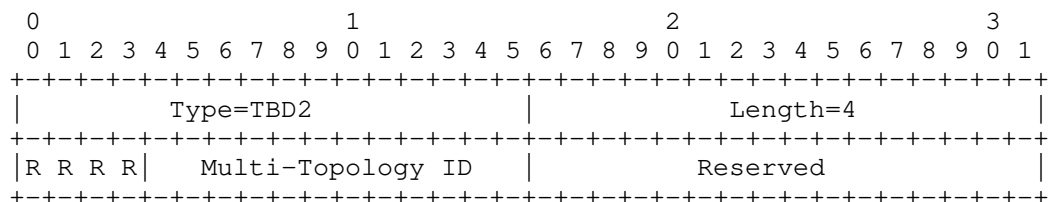


Figure 2: Multi-topology TLV

The code point for the TLV type is TBD2. The TLV length is 4 octets.

Multi-Topology ID (12 bits): Semantics of the IS-IS MT-ID are defined in Section 7.2 of [RFC5120]. Semantics of the OSPF MT-ID are defined in Section 3.7 of [RFC4915]. If the value is derived from OSPF, then the upper 9 bits MUST be set to 0. Bits R are reserved and SHOULD be set to 0 when originated and ignored on receipt.

Reserved (16 bits): This field MUST be set to zero on transmission and MUST be ignored on receipt.

3.3. Slice-id TLV

PCEP message needs to carry Slice ID to let the scope of path calculation to be limited in a specific slice.

There are many control plane technologies to realize slicing. Some control plane technologies may directly maintain resources per slice granularity in the link-state database, only for the case with small slice scalability. [I-D.bestbar-teas-ns-packet] proposes a more scalable slicing scheme. The resource information in link-state database is identified by SA-ID to distinguish the logical topologies corresponding to different slice-aggregate. Within the controller, a slice-aggregate includes one or more slices mapped to it. If the number of slices is small, the resources per slice granularity can be maintained directly in the link-state database. In this case, different slice may be mapped to different slice-aggregate. If the number of slices is large, it is not recommended to maintain the slice granularity resources in the link-state database, but the aggregated SA-ID granularity.

In any case, the slice service (such as VPN service) perceives the Slice ID (not others), so it is natural for the service to include a Slice ID constraint in its TE purpose definition. For example, VPN routes may have Color attribute (refer to [I-D.ietf-idr-tunnel-encaps] and [I-D.ietf-spring-segment-routing-policy]). Color represents a

specific TE purpose, which can contain a Slice ID. Thus it is natural carry Slice ID in PCEP message.

When the controller receives the path computation request with a Slice ID constraint, it can use the resources identified by specific Slice in TED, or firstly look up the Slice ID to SA-ID mapping entry and then use the resources of specific SA-ID in TED, to calculate the path.

In a PCReq message, a PCC MAY insert one Slice-id TLV to indicate the slice based virtual network that MUST be considered by the PCE. The PCE will perform path computation based on the intra-domain or inter-domain sub-topology identified by the specific Slice-id, which is independent of routing protocols such as IGP/BGP. The absence of the Slice-id TLV MUST be interpreted by the PCE as a path computation request for which no constraints need be applied to any of slice, i.e, a default Slice-id (0) will be applied.

In a PCRep/PCInit/PCUpd message, the Slice-id TLV MAY be carried so as to provide the network slicing information for the computed path. The headend may put the Slice-id to an encapsulated data packet.

The format of the Slice-id TLV is shown as Figure 3:

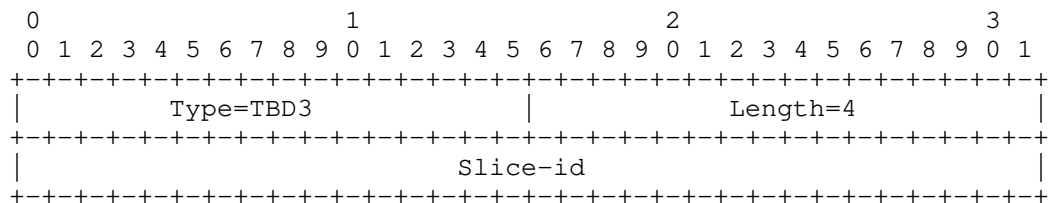


Figure 3: Slice-id TLV

The code point for the TLV type is TBD3. The TLV length is 4 octets.

Slice-id (32 bits): indicate the Slice-id information. The Slice-id is also termed as AII defined in [I-D.peng-lsr-network-slicing] to represent an IETF Network Slice that is defined in [I-D.ietf-teas-ietf-network-slice-definition].

3.4. Application Specific TLV

The Application Specific TLV is optional and is defined to carry the application specific constraints.

In a PCReq message, a PCC MAY insert one Application Specific TLV to indicate the application that MUST be considered by the PCE. The PCE will perform path computation using the specific application attributes. The absence of the Application Specific TLV MUST be interpreted by the PCE as a path computation request for which no constraints need be applied to any of the Application Specific attributes.

In a PCRep/PCInit/PCUpd message, the Application Specific TLV MAY be inserted so as to provide the Application Specific information for the computed path.

The format of the Application Specific TLV is shown as Figure 4:

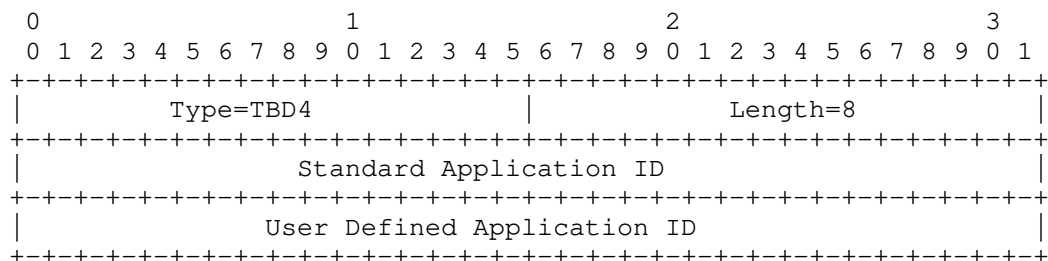


Figure 4: Application Specific TLV

The code point for the TLV type is TBD4. The TLV length is 8 octets.

Standard Application ID: Represents a bit-position value for a single STANDARD application that is defined in the IANA "IGP Parameters" registries under the "Link Attribute Applications" registry [RFC8919].

User Defined Application ID: Represents a single user defined application which is a specific implementation.

3.5. Color TLV

The Color TLV is optional and is defined to carry the color constraints.

In a PCReq message, a PCC MAY insert one Color TLV to indicate the traffic engineering purpose that is recognized by both PCE and PCC with no conflict meaning. The PCE will perform path computation based on the color template. The same color template may be also defined at PCC and the existing constraints (i.e, metric, bandwidth, delay, etc) carried in the message MUST be ignored. The absence of

the Color TLV MUST be interpreted by the PCE as a path computation request for which traditional constraints that are contained in message need be applied.

In a PCRep/PCInit/PCUpd message, the Color TLV MAY be inserted so as to provide the TE purpose information for the computed path, the PCC recognize the color value that match a local color-template. For example, the COLOR TLV can be used to identify the Color of each Candidate Path in the Composite Candidate Path as described in [I-D.ietf-pce-multipath]

The format of the Color TLV is shown as Figure 5:

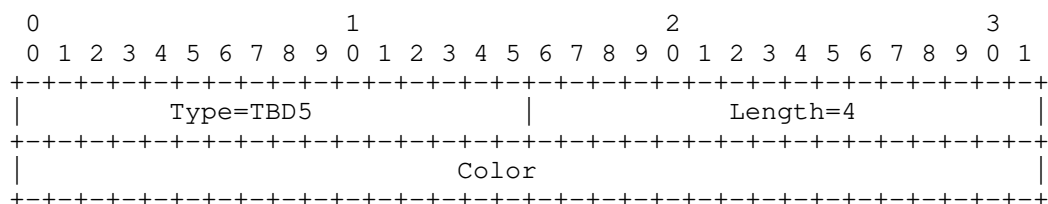


Figure 5: Color TLV

The code point for the TLV type is TBD5. The TLV length is 4 octets.

Color (32 bits): indicate a TE template, 0 is invalid value. It is consistent with the Color Extended Community defined in [I-D.ietf-idr-tunnel-encaps], and color of SR policy defined in [I-D.ietf-spring-segment-routing-policy].

Note that Color TLV defined in this document is used to represent a TE template, it can be suitable for any TE instance such as RSVP-TE, SR-TE, SR-policy. [I-D.ietf-pce-segment-routing-policy-cp] has proposed the SR policy KEY (that also includes a color information) as an association group KEY to associate many candidate paths, however it is only for association purpose but not constraint purpose for path computation.

A color template can be defined to contain existing constraints such as metric, bandwidth, delay, affinity parameters, and the sub-topology constraints above defined in this document.

3.6. FA-id TLV

FA-id defined in [I-D.ietf-lsr-flex-algo] is a short mapping of SR policy color to optimize segment stack depth for the IGP area partial of the entire SR policy. The overlay service that want to be carried over a particular SR-FA path must firstly let the SR policy supplier know that requirement. There are two possible ways to map a color to an FA-id. One is explicit mapping configuration within color template, the other is dynamically replacing a long segment list to short FA segment by headend or controller once the constraints contained in the color-template equal to that contained in FAD.

In addition to the above mapping behavior, it is also possible to merge the constraints contained in the color-template and constraints contained in FAD. The merging behavior can be used to compute SR-TE path within a Flex-algo plane.

In a PCReq message, a PCC MAY insert one FA-id TLV to indicate the above explicit FA-id mapping or merging. For mapping case, the PCE will perform path computation based on the FA-id mapping. In detailed, The PCE will check if there are connectivity within the corresponding Flex-algo plane to the destination. If yes, the path computation result will be represented as segment list with a single prefix-SID@FA for intra-domain case, or several prefix-SID@FA for inter-domain case.

For merging case, the PCE will perform path computation based on the total constraints combined with the ones contained in FAD identified by FA-id and other ones contained in PCReq message. The later constraints can get from color template or directly represent by a color. In this case the computed path will be limited in the specific Flex-algo plane determined by link resource Including/ Excluding rules of FAD, and at the same time the path will also meet other constraints for the TE purpose within the Flex-algo plane. The PCE can optimize the strictly path to a loosely path when a part of the strictly path is consistent with the algorithm based path, i.e, some consecutive adjacency SIDs can be replaced with a single algorithm based Prefix-SID.

In a PCRep/PCInit/PCUpd message, the FA-id TLV MAY be inserted so as to provide the FA plane information for the computed path.

In general, the FA-id TLV is only meaningful for the domain (ingress domain) that headend node belongs to. For inter-domain case, operator SHOULD ensure the FA-id configuration of different domain are same for an E2E slice, when he want to explicitly indicate FA-id in PCEP message, otherwise the PCE has to choose different FA-id for

other domain as long as the contents of FAD is consistent with the one of ingress domain.

The format of the FA-id TLV is shown as Figure 6:

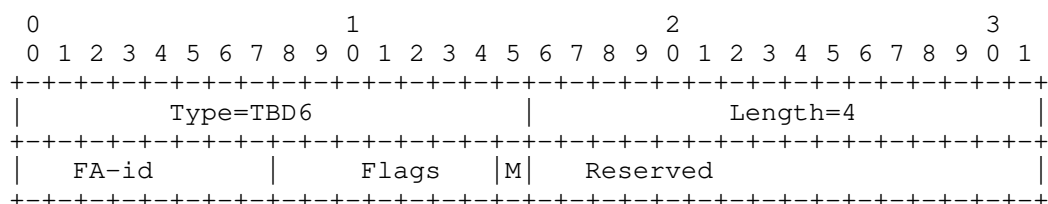


Figure 6: FA-id TLV

The code point for the TLV type is TBD6. The TLV length is 4 octets.

FA-id (8 bits): indicate an explicit FA-id mapping information.

Flags (8 bits): Currently only one flag, Flag-M, is defined.

Flag-M: Indicate mapping behavior when unset, and merging behavior when set.

4. Security Considerations

TBA

5. Acknowledgements

TBA

6. IANA Considerations

IANA is requested to make allocations from the registry, as follows:

Type	TLV	Reference
TBD1	Source Protocol TLV	[this document]
TBD2	Multi-topology TLV	[this document]
TBD3	Slice-id TLV	[this document]
TBD4	Application Specific TLV	[this document]
TBD5	Color TLV	[this document]
TBD6	FA-id TLV	[this document]

Table 1

7. Normative References

[I-D.bestbar-teas-ns-packet]

Saad, T., Beeram, V. P., Wen, B., Ceccarelli, D., Halpern, J., Peng, S., Chen, R., Liu, X., and L. M. Contreras, "Realizing Network Slices in IP/MPLS Networks", draft-bestbar-teas-ns-packet-02 (work in progress), February 2021.

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

Patel, K., Velde, G. V. D., Sangli, S. R., and J. Scudder, "The BGP Tunnel Encapsulation Attribute", draft-ietf-idr-tunnel-encaps-22 (work in progress), January 2021.

[I-D.ietf-lsr-flex-algo]

Psenak, P., Hegde, S., Filsfils, C., Talaulikar, K., and A. Gulko, "IGP Flexible Algorithm", draft-ietf-lsr-flex-algo-15 (work in progress), April 2021.

[I-D.ietf-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-ietf-pce-multipath-00 (work in progress), May 2021.

[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-04 (work in progress), March 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-11 (work in progress), April 2021.
- [I-D.ietf-teas-ietf-network-slice-definition]
Rokui, R., Homma, S., Makhiyani, K., Contreras, L. M., and J. Tantsura, "Definition of IETF Network Slices", draft-ietf-teas-ietf-network-slice-definition-01 (work in progress), February 2021.
- [I-D.peng-lsr-network-slicing]
Peng, S., Chen, R., and G. Mirsky, "Packet Network Slicing using Segment Routing", draft-peng-lsr-network-slicing-00 (work in progress), February 2019.
- [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>>.
- [RFC4915] Psenak, P., Mirtorabi, S., Roy, A., Nguyen, L., and P. Pillay-Esnault, "Multi-Topology (MT) Routing in OSPF", RFC 4915, DOI 10.17487/RFC4915, June 2007, <<https://www.rfc-editor.org/info/rfc4915>>.
- [RFC5120] Przygienda, T., Shen, N., and N. Sheth, "M-ISIS: Multi Topology (MT) Routing in Intermediate System to Intermediate Systems (IS-IS)", RFC 5120, DOI 10.17487/RFC5120, February 2008, <<https://www.rfc-editor.org/info/rfc5120>>.
- [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>>.
- [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>>.

- [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>>.
- [RFC8919] Ginsberg, L., Psenak, P., Previdi, S., Henderickx, W., and J. Drake, "IS-IS Application-Specific Link Attributes", RFC 8919, DOI 10.17487/RFC8919, October 2020, <<https://www.rfc-editor.org/info/rfc8919>>.

Authors' Addresses

Shaofu Peng
ZTE Corporation
No.50 Software Avenue
Nanjing, Jiangsu 210012
China

Email: peng.shaofu@zte.com.cn

Quan Xiong
ZTE Corporation
No.6 Huashi Park Rd
Wuhan, Hubei 430223
China

Email: xiong.quan@zte.com.cn

Fengwei Qin
China Mobile
Beijing
China

Email: qinfengwei@chinamobile.com

Mike Koldychev
Cisco Systems
Canada

Email: mkoldych@cisco.com

Siva Sivabalan
Ciena Corporation
Canada

Email: ssivabal@ciena.com

PCE Working Group
Internet-Draft
Intended status: Standards Track
Expires: January 13, 2022

B. Rajagopalan
V. Beeram
Juniper Networks
G. Mishra
Verizon Communications Inc.
July 12, 2021

Path Computation Element Protocol (PCEP) Extension for RSVP Color
draft-rajagopalan-pcep-rsvp-color-01

Abstract

This document specifies extensions to Path Computation Element Protocol (PCEP) to carry a newly defined attribute of RSVP LSP called 'color' that can be used as a guiding criterion for selecting the LSP as a next hop for a service route.

Requirements Language

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

Status of This Memo

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

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at <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 January 13, 2022.

Copyright Notice

Copyright (c) 2021 IETF Trust and the persons identified as the document authors. All rights reserved.

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

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

Table of Contents

1. Introduction	2
2. Protocol Operation	3
3. TLV Format	4
4. Usage with BGP-CT	4
5. Security Considerations	4
6. IANA Considerations	5
7. Acknowledgments	5
8. References	5
8.1. Normative References	5
8.2. Informative References	6
Authors' Addresses	6

1. Introduction

This document defines a new RSVP LSP property, called "color", that can be exchanged over PCEP. The 'color' field can be used as one of the guiding criteria in selecting the LSP as a next hop for service prefixes.

While the specific details of how the service prefixes are associated with the appropriate RSVP LSP's are outside the scope of this specification, the envisioned high level usage of the 'color' field is as follows.

The service prefixes are marked with some indication of the type of underlay they need. The underlay LSP's carry corresponding markings, which we refer to as "color" in this specification, enabling an ingress node to associate the service prefixes with the appropriate underlay LSP's.

As an example, for a BGP-based service, the originating PE could attach some community, e.g. the Extended Color Community [RFC5512] with the service route. A receiving PE could use locally configured policies to associate service routes carrying Extended Color Community 'X' with underlay RSVP LSP's of color 'Y'.

While the Extended Color Community provides a convenient method to perform the mapping, the policy on the ingress node is free to

classify on any property of the route to select underlay RSVP LSP's of a certain color.

The 'color' specified in this draft is mainly used for facilitating underlay selection, and does not have any effect on the constraints used for path computation.

2. Protocol Operation

The STATEFUL-PCE-CAPABILITY negotiation message is enhanced to carry the color capability, which allows PCC & PCE to determine how incompatibility should be handled, should only one of them support color. An older implementation that does not recognize the new color TLV would ignore it upon receipt. This can sometimes result in undesirable behavior. For example, if PCE passes color to a PCC that does not understand colors, the LSP may not be used as intended. A PCE that clearly knows the PCC's color capability can handle such cases better, and vice versa. Following are the rules for handling mismatch in color capability.

A PCE that has color capability MUST NOT send color TLV to a PCC that does not have color capability. A PCE that does not have color capability can ignore color marking reported by PCC.

When a PCC is interacting with a PCE that does not have color capability, the PCC

- o SHOULD NOT report color to the PCE.
- o MUST NOT override the local color, if it is configured, based on any messages coming from the PCE.

The actual color value itself is carried in a newly defined TLV in the LSP Object defined in [RFC8231].

If a PCC is unable to honor a color value passed in an LSP Update request, the PCC must keep the LSP in DOWN state, and include an LSP Error Code value of "Unsupported Color" [Value to be assigned by IANA] in LSP State Report message.

If an RSVP tunnel has multiple LSP's associated with it, the PCE should designate one of the LSP's as primary, and attach the color with that LSP. If PCC receives color TLV for an LSP that it treats as secondary, it SHOULD respond with an error code of 4 (Unacceptable Parameters).

3. TLV Format

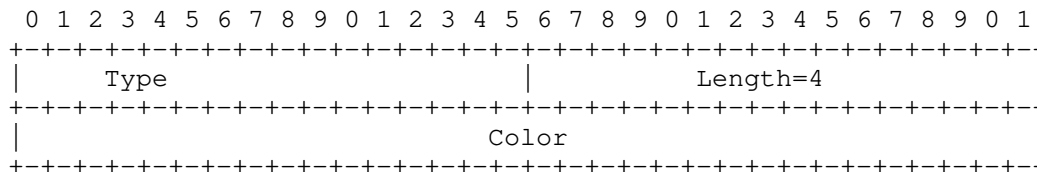


Figure 1: Color TLV in LSP Object

Type has the value [TO-BE-ASSIGNED-BY-IANA]. Length carries a value of 4. The 'color' field is 4-bytes long, and carries the actual color value.

Section 7.1.1 of RFC8231 [RFC8231] defines STATEFUL-PCE-CAPABILITY flags. The following flag is used to indicate if the speaker supports color capability:

C-bit (TO-BE-ASSIGNED-BY-IANA): A PCE/PCC that supports color capability must turn on this bit.

4. Usage with BGP-CT

RSVP LSP's marked with color can also be used for inter-domain service mapping as defined in BGP-CT [I-D.kaliraj-idr-bgp-classful-transport-planes]. In BGP-CT, the mapping community of the service route is used to select a "resolution scheme", which in turn selects LSP's of various "transport classes" in the defined order of preference. The 'color' field defined in this specification could be used to associate the RSVP LSP with a particular transport class.

A colored RSVP LSP can also be exported into BGP-CT for inter-domain classful transport.

5. Security Considerations

This document defines a new TLV for color, and a new flag in capability negotiation, which do not add any new security concerns beyond those discussed in [RFC5440], [RFC8231] and [RFC8281].

An unauthorized PCE may maliciously associate the LSP with an incorrect color. The procedures described in [RFC8253] and [RFC7525] can be used to protect against this attack.

6. IANA Considerations

IANA is requested to assign code points for the following:

- o Code point for "Color" TLV from the sub-registry "PCEP TLV Type Indicators".
- o C-bit value from the sub-registry "STATEFUL-PCE-CAPABILITY TLV Flag Field".
- o An error code for "Unsupported color" from the sub-registry "LSP-ERROR-CODE TLV Error Code Field".

7. Acknowledgments

The authors would like to thank Kaliraj Vairavakkalai, Colby Barth & Natrajan Venkataraman for their review & suggestions, which helped improve this specification.

8. References

8.1. Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <<https://www.rfc-editor.org/info/rfc2119>>.
- [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>>.
- [RFC5512] Mohapatra, P. and E. Rosen, "The BGP Encapsulation Subsequent Address Family Identifier (SAFI) and the BGP Tunnel Encapsulation Attribute", RFC 5512, DOI 10.17487/RFC5512, April 2009, <<https://www.rfc-editor.org/info/rfc5512>>.
- [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>>.

- [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>>.

8.2. Informative References

- [I-D.kaliraj-idr-bgp-classful-transport-planes]
Vairavakkalai, K., Venkataraman, N., Rajagopalan, B., Mishra, G., Khaddam, M., Xu, X., and R. J. Szarecki, "BGP Classful Transport Planes", draft-kaliraj-idr-bgp-classful-transport-planes-07 (work in progress), February 2021.

Authors' Addresses

Balaji Rajagopalan
Juniper Networks

Email: balajir@juniper.net

Vishnu Pavan Beeram
Juniper Networks

Email: vbeeram@juniper.net

Gyan Mishra
Verizon Communications Inc.

Email: gyan.s.mishra@verizon.com

PCE Working Group
Internet-Draft
Intended status: Standards Track
Expires: January 13, 2022

A. Tokar
S. Sidor
Cisco Systems, Inc.
S. Sivabalan
Ciena
S. Peng
Huawei Technologies
M. Negi
RtBrick Inc
July 12, 2021

Carrying SID Algorithm information in PCE-based Networks.
draft-tokar-pce-sid-algo-04

Abstract

The Algorithm associated with a prefix Segment-ID (SID) defines the path computation Algorithm used by Interior Gateway Protocols (IGPs). This information is available to controllers such as the Path Computation Element (PCE) via topology learning. This document proposes an approach for informing headend routers regarding the Algorithm associated with each prefix SID used in PCE-computed paths, as well as signalling a specific SID algorithm as a constraint to the PCE.

Requirements Language

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

Status of This Memo

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

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

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

This Internet-Draft will expire on January 13, 2022.

Copyright Notice

Copyright (c) 2021 IETF Trust and the persons identified as the document authors. All rights reserved.

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

Table of Contents

1. Introduction	2
2. Terminology	3
3. Object Formats	4
3.1. OPEN Object	4
3.1.1. SR PCE Capability Sub-TLV	4
3.1.2. SRv6 PCE Capability sub-TLV	4
3.2. SR-ERO Subobject	5
3.3. SRv6-ERO Subobject	5
3.4. LSPA Object	5
4. Operation	6
4.1. SR-ERO and SRv6-ERO Encoding	6
4.2. SID Algorithm Constraint	7
5. Security Considerations	7
6. IANA Considerations	7
6.1. SR Capability Flag	7
6.2. SRv6 PCE Capability Flag	7
6.3. SR-ERO Flag	8
6.4. SRv6-ERO Flag	8
6.5. PCEP TLV Types	8
7. Normative References	8
Appendix A. Contributors	9
Authors' Addresses	9

1. Introduction

A PCE can compute SR-TE paths using SIDs with different Algorithms depending on the use-case, constraints, etc. While this information is available on the PCE, there is no method of conveying this information to the headend router.

Similarly, the headend can also compute SR-TE paths using different Algorithms, and this information also needs to be conveyed to the PCE for collection or troubleshooting purposes. In addition, in the case of multiple (redundant) PCEs, when the headend receives a path from the primary PCE, it needs to be able to report the complete path information - including the Algorithm - to the backup PCE so that in HA scenarios, the backup PCE can verify the prefix SIDs appropriately.

An operator may also want to constrain the path computed by the PCE to a specific SID Algorithm, for example, in order to only use SID Algorithms for a low-latency path. A new TLV is introduced for this purpose.

Refer to [RFC8665] and [RFC8667] for details about the prefix SID Algorithm.

This document is extending:

- o the SR PCE Capability Sub-TLV and the SR-ERO subobject - defined in [RFC8664]
- o the SRv6 PCE Capability sub-TLV and the SRv6-ERO subobject - defined in [I-D.ietf-pce-segment-routing-ipv6]

A new TLV for signalling SID Algorithm constraint to the PCE is also introduced, to be carried inside the LSPA object, which is defined in [RFC5440].

The mechanisms described in this document are equally applicable to both SR-MPLS and SRv6.

2. Terminology

The following terminologies are used in this document:

ERO: Explicit Route Object

IGP: Interior Gateway Protocol

NAI: Node or Adjacency Identifier.

PCE: Path Computation Element

PCEP: Path Computation Element Protocol.

SID: Segment Identifier.

SR: Segment Routing.

SR-TE: Segment Routing Traffic Engineering.

LSP: Label Switched Path.

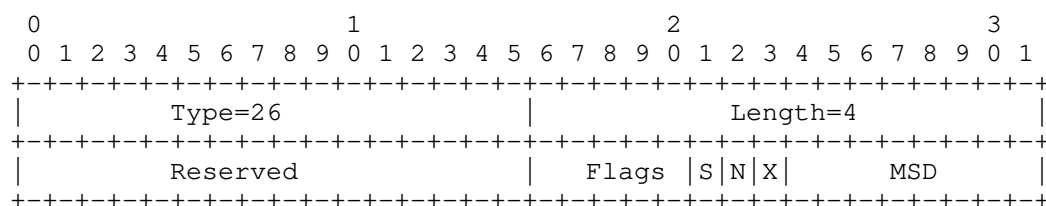
LSPA: Label Switched Path Attributes.

3. Object Formats

3.1. OPEN Object

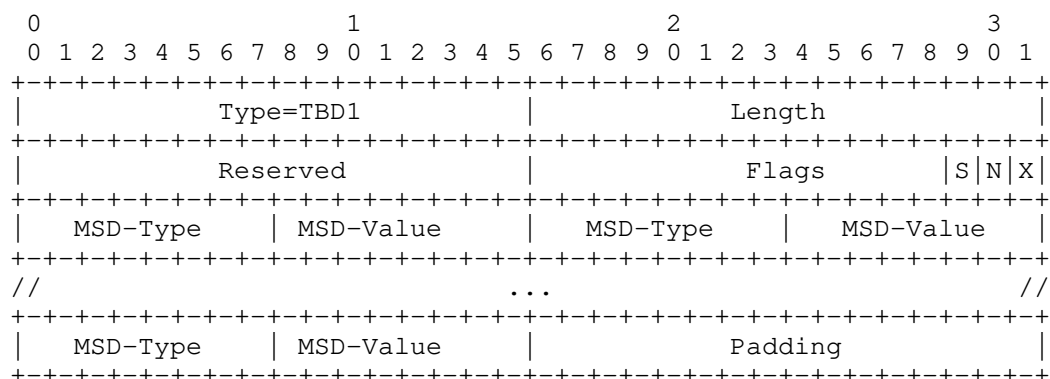
3.1.1. SR PCE Capability Sub-TLV

A new flag S is proposed in the SR PCE Capability Sub-TLV introduced in Section 4.1.2 of [RFC8664] in Path Computation Element Communication Protocol (PCEP) to indicate support for SID Algorithm field in the SR-ERO subobject.



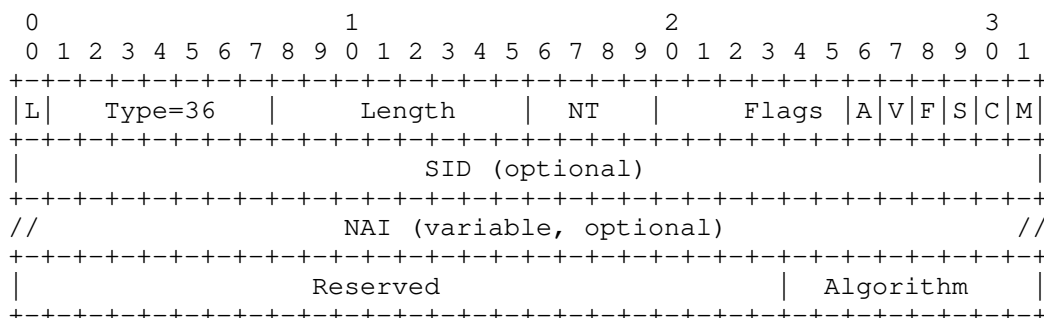
3.1.2. SRv6 PCE Capability sub-TLV

A new flag S is proposed in the SRv6 PCE Capability sub-TLV introduced in 4.1.1 of [I-D.ietf-pce-segment-routing-ipv6] in Path Computation Element Communication Protocol (PCEP) to indicate support for SID Algorithm field in the SRv6-ERO subobject.



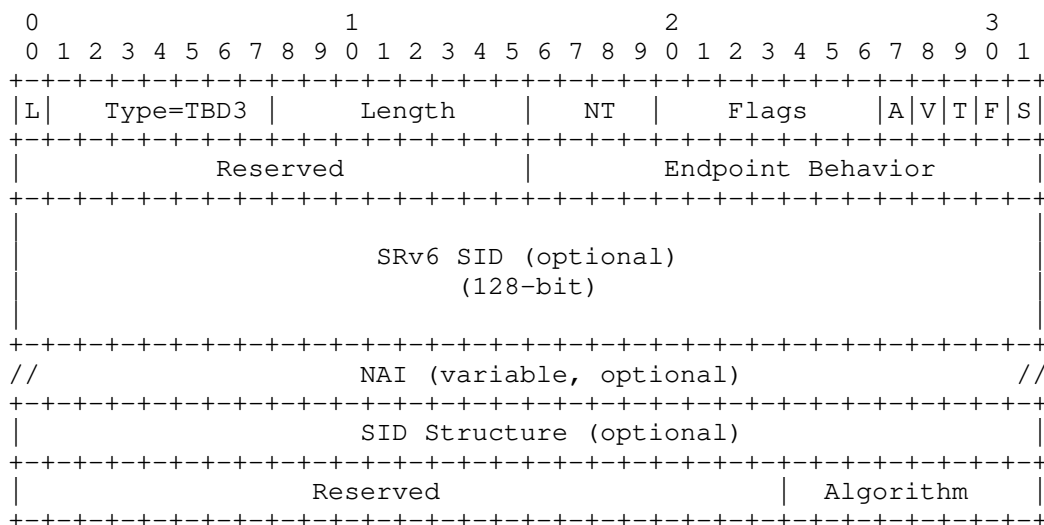
3.2. SR-ERO Subobject

The SR-ERO subobject encoding is extended with new flag "A" to indicate if the Algorithm field is included after other optional fields.



3.3. SRv6-ERO Subobject

The SRv6-ERO subobject encoding is extended with new flag "A" to indicate if the Algorithm field is included after other optional fields.



3.4. LSPA Object

A new TLV for the LSPA Object with TLV type=TBD3 is introduced to carry the SID Algorithm constraint. This TLV SHOULD only be used when PST (Path Setup type) = SR or SRv6.

The format of the SID Algorithm TLV is as follows:

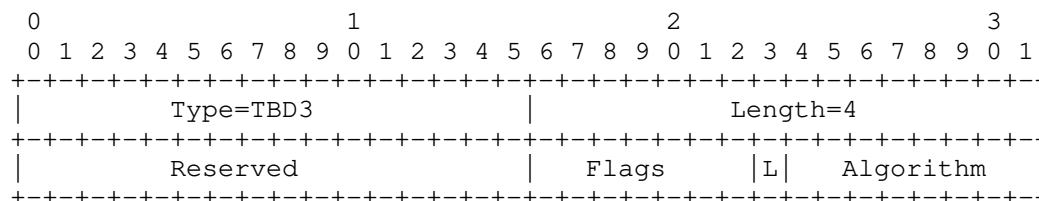


Figure 1: SID Algorithm TLV Format

The code point for the TLV type is TBD3. The TLV length is 4 octets.

The 32-bit value is formatted as follows.

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.

- * **L (Loose):** If set to 1, the PCE MAY insert SIDs with a different Algorithm, but it MUST prefer the specified Algorithm whenever possible.

Algorithm: SID Algorithm the PCE MUST take into account while computing a path for the LSP.

4. Operation

4.1. SR-ERO and SRv6-ERO Encoding

PCEP speaker MAY set the A flag and include the Algorithm field in SR-ERO or SRv6-ERO subobject if the S flag was advertised by both PCEP speakers.

If PCEP peer receives SR-ERO subobject with the A flag set or with the SID Algorithm included, but the S flag was not advertised, then such PCEP message must be rejected with PCErrror as described in Section 7.2 of [RFC5440]

The Algorithm field MUST be included after optional SID, NAI or SID structure and length of SR-ERO or SRv6-ERO subobject MUST be increased with additional 4 bytes for Reserved and Algorithm field.

4.2. SID Algorithm Constraint

In order to signal a specific SID Algorithm constraint to the PCE, the headend MUST encode the SID ALGORITHM TLV inside the LSPA object.

When the PCE receives a SID Algorithm constraint, it MUST only take prefix SIDs with the specified Algorithm into account during path computation. However, if the L flag is set in the SID Algorithm TLV, the PCE MAY insert prefix SIDs with a different Algorithm in order to successfully compute a path.

If the PCE is unable to find a path with the given SID Algorithm constraint, it MUST bring the LSP down.

SID Algorithm does not replace the Objective Function defined in [RFC5541]. The SID Algorithm constraint acts as a filter, restricting which SIDs may be used as a result of the path computation function.

5. Security Considerations

No additional security measure is required.

6. IANA Considerations

6.1. SR Capability Flag

IANA maintains a sub-registry, named "SR Capability Flag Field", within the "Path Computation Element Protocol (PCEP) Numbers" registry to manage the Flags field of the SR-PCE-CAPABILITY TLV. IANA is requested to make the following assignment:

Value	Description	Reference
TBD1	SID Algorithm Capability	This document

6.2. SRv6 PCE Capability Flag

IANA was requested in [I-D.ietf-pce-segment-routing-ipv6] to create a sub-registry, named "SRv6 PCE Capability Flags", within the "Path Computation Element Protocol (PCEP) Numbers" registry to manage the Flags field of SRv6-PCE-CAPABILITY sub-TLV. IANA is requested to make the following assignment:

Value	Description	Reference
TBD2	SID Algorithm Capability	This document

6.3. SR-ERO Flag

IANA maintains a sub-registry, named "SR-ERO Flag Field", within the "Path Computation Element Protocol (PCEP) Numbers" registry to manage the Flags field of the SR-ERO Subobject. IANA is requested to make the following assignment:

Value	Description	Reference
TBD3	SID Algorithm Flag	This document

6.4. SRv6-ERO Flag

IANA was requested in [I-D.ietf-pce-segment-routing-ipv6], named "SRv6-ERO Flag Field", within the "Path Computation Element Protocol (PCEP) Numbers" registry to manage the Flags field of the SRv6-ERO subobject. IANA is requested to make the following assignment:

Value	Description	Reference
TBD4	SID Algorithm Flag	This document

6.5. PCEP TLV Types

IANA is requested to allocate a new TLV type for the new LSPA TLV specified in this document.

Value	Description	Reference
TBD5	SID Algorithm	This document

7. Normative 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-ipv6-09 (work in progress), May 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>>.
- [RFC5541] Le Roux, JL., Vasseur, JP., and Y. Lee, "Encoding of Objective Functions in the Path Computation Element Communication Protocol (PCEP)", RFC 5541, DOI 10.17487/RFC5541, June 2009, <<https://www.rfc-editor.org/info/rfc5541>>.
- [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>>.
- [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>>.

Appendix A. Contributors

Mike Koldychev
Cisco Systems
Kanata, Ontario
Canada

Email: mkoldych@cisco.com

Authors' Addresses

Alex Tokar
Cisco Systems, Inc.
Eurovea Central 3.
Pribinova 10
Bratislava 811 09
Slovakia

Email: atokar@cisco.com

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

Email: ssidor@cisco.com

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

Email: msiva282@gmail.com

Shuping Peng
Huawei Technologies
Huawei Campus, No. 156 Beiqing Rd.
Beijing 100095
China

Email: pengshuping@huawei.com

Mahendra Singh Negi
RtBrick Inc
Bangalore, Karnataka
India

Email: mahend.ietf@gmail.com

PCE Working Group
Internet-Draft
Intended status: Standards Track
Expires: 7 April 2022

A. Tokar
S. Sidor
Cisco Systems, Inc.
S. Peng
ZTE Corporation
S. Sivabalan
Ciena
T. Saad
Juniper Networks
S. Peng
Huawei Technologies
M. Negi
RtBrick Inc
4 October 2021

Carrying SID Algorithm information in PCE-based Networks.
draft-tokar-pce-sid-algo-05

Abstract

The Algorithm associated with a prefix Segment-ID (SID) defines the path computation Algorithm used by Interior Gateway Protocols (IGPs). This information is available to controllers such as the Path Computation Element (PCE) via topology learning. This document proposes an approach for informing headend routers regarding the Algorithm associated with each prefix SID used in PCE-computed paths, as well as signalling a specific SID algorithm as a constraint to the PCE.

Requirements Language

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

Status of This Memo

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

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

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

This Internet-Draft will expire on 7 April 2022.

Copyright Notice

Copyright (c) 2021 IETF Trust and the persons identified as the document authors. All rights reserved.

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

Table of Contents

1. Introduction	3
2. Terminology	3
3. Object Formats	4
3.1. OPEN Object	4
3.1.1. SR PCE Capability Sub-TLV	4
3.1.2. SRv6 PCE Capability sub-TLV	4
3.2. SR-ERO Subobject	5
3.3. SRv6-ERO Subobject	5
3.4. LSPA Object	6
4. Operation	7
4.1. SR-ERO and SRv6-ERO Encoding	7
4.2. SID Algorithm Constraint	7
5. Security Considerations	7
6. IANA Considerations	8
6.1. SR Capability Flag	8
6.2. SRv6 PCE Capability Flag	8
6.3. SR-ERO Flag	8
6.4. SRv6-ERO Flag	9
6.5. PCEP TLV Types	9
7. Normative References	9
Appendix A. Contributors	10
Authors' Addresses	11

1. Introduction

A PCE can compute SR-TE paths using SIDs with different Algorithms depending on the use-case, constraints, etc. While this information is available on the PCE, there is no method of conveying this information to the headend router.

Similarly, the headend can also compute SR-TE paths using different Algorithms, and this information also needs to be conveyed to the PCE for collection or troubleshooting purposes. In addition, in the case of multiple (redundant) PCEs, when the headend receives a path from the primary PCE, it needs to be able to report the complete path information - including the Algorithm - to the backup PCE so that in HA scenarios, the backup PCE can verify the prefix SIDs appropriately.

An operator may also want to constrain the path computed by the PCE to a specific SID Algorithm, for example, in order to only use SID Algorithms for a low-latency path. A new TLV is introduced for this purpose.

Refer to [RFC8665] and [RFC8667] for details about the prefix SID Algorithm.

This document is extending:

- * the SR PCE Capability Sub-TLV and the SR-ERO subobject - defined in [RFC8664]
- * the SRv6 PCE Capability sub-TLV and the SRv6-ERO subobject - defined in [I-D.ietf-pce-segment-routing-ipv6]

A new TLV for signalling SID Algorithm constraint to the PCE is also introduced, to be carried inside the LSPA object, which is defined in [RFC5440].

The mechanisms described in this document are equally applicable to both SR-MPLS and SRv6.

2. Terminology

The following terminologies are used in this document:

ERO: Explicit Route Object

IGP: Interior Gateway Protocol

NAI: Node or Adjacency Identifier.

PCE: Path Computation Element

PCEP: Path Computation Element Protocol.

SID: Segment Identifier.

SR: Segment Routing.

SR-TE: Segment Routing Traffic Engineering.

LSP: Label Switched Path.

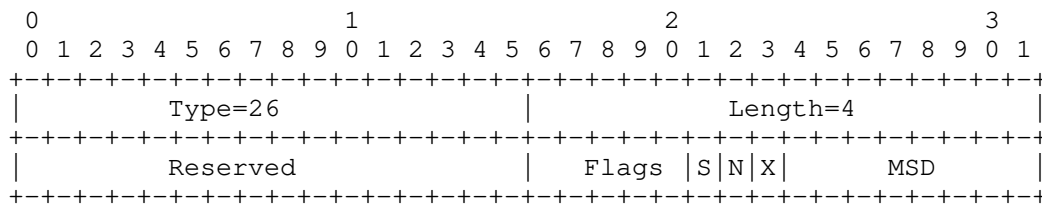
LSPA: Label Switched Path Attributes.

3. Object Formats

3.1. OPEN Object

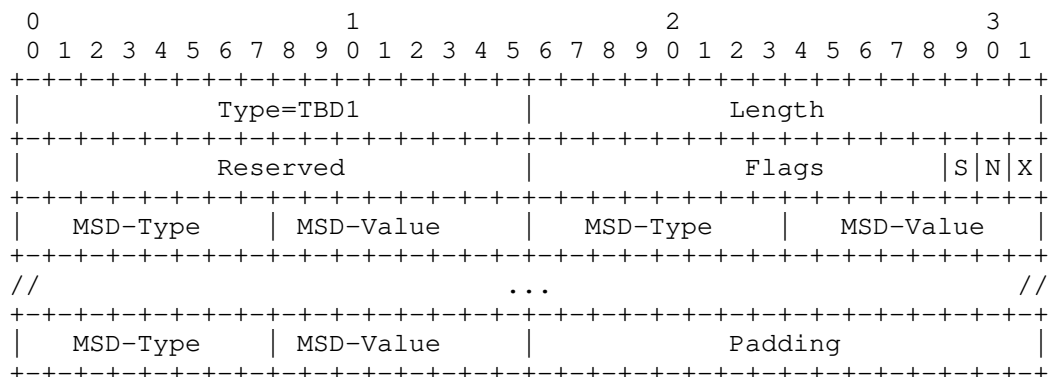
3.1.1. SR PCE Capability Sub-TLV

A new flag S is proposed in the SR PCE Capability Sub-TLV introduced in Section 4.1.2 of [RFC8664] in Path Computation Element Communication Protocol (PCEP) to indicate support for SID Algorithm field in the SR-ERO subobject.



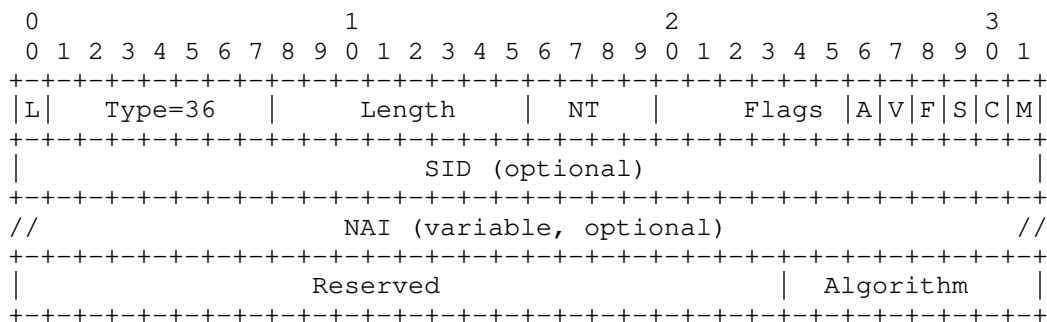
3.1.2. SRv6 PCE Capability sub-TLV

A new flag S is proposed in the SRv6 PCE Capability sub-TLV introduced in 4.1.1 of [I-D.ietf-pce-segment-routing-ipv6] in Path Computation Element Communication Protocol (PCEP) to indicate support for SID Algorithm field in the SRv6-ERO subobject.



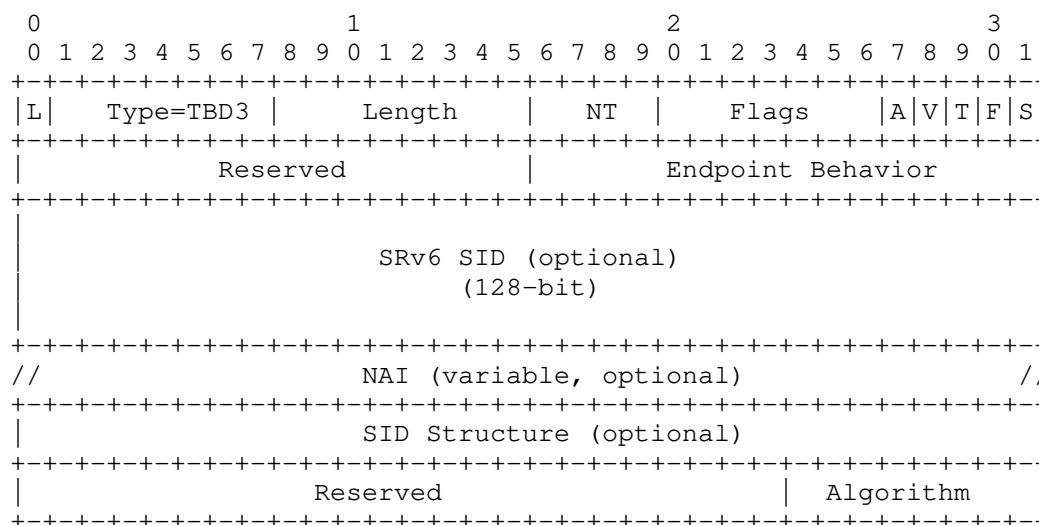
3.2. SR-ERO Subobject

The SR-ERO subobject encoding is extended with new flag "A" to indicate if the Algorithm field is included after other optional fields.



3.3. SRv6-ERO Subobject

The SRv6-ERO subobject encoding is extended with new flag "A" to indicate if the Algorithm field is included after other optional fields.



3.4. LSPA Object

A new TLV for the LSPA Object with TLV type=TBD3 is introduced to carry the SID Algorithm constraint. This TLV SHOULD only be used when PST (Path Setup type) = SR or SRv6.

The format of the SID Algorithm TLV is as follows:

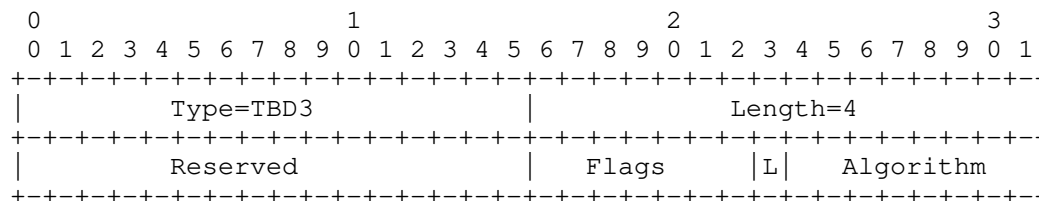


Figure 1: SID Algorithm TLV Format

The code point for the TLV type is TBD3. The TLV length is 4 octets.

The 32-bit value is formatted as follows.

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.

- * L (Loose): If set to 1, the PCE MAY insert SIDs with a different Algorithm, but it MUST prefer the specified Algorithm whenever possible.

Algorithm: SID Algorithm the PCE MUST take into account while computing a path for the LSP.

4. Operation

4.1. SR-ERO and SRv6-ERO Encoding

PCEP speaker MAY set the A flag and include the Algorithm field in SR-ERO or SRv6-ERO subobject if the S flag was advertised by both PCEP speakers.

If PCEP peer receives SR-ERO subobject with the A flag set or with the SID Algorithm included, but the S flag was not advertised, then such PCEP message must be rejected with PCErrror as described in Section 7.2 of [RFC5440]

The Algorithm field MUST be included after optional SID, NAI or SID structure and length of SR-ERO or SRv6-ERO subobject MUST be increased with additional 4 bytes for Reserved and Algorithm field.

4.2. SID Algorithm Constraint

In order to signal a specific SID Algorithm constraint to the PCE, the headend MUST encode the SID ALGORITHM TLV inside the LSPA object.

When the PCE receives a SID Algorithm constraint, it MUST only take prefix SIDs with the specified Algorithm into account during path computation. However, if the L flag is set in the SID Algorithm TLV, the PCE MAY insert prefix SIDs with a different Algorithm in order to successfully compute a path.

If the PCE is unable to find a path with the given SID Algorithm constraint, it MUST bring the LSP down.

SID Algorithm does not replace the Objective Function defined in [RFC5541]. The SID Algorithm constraint acts as a filter, restricting which SIDs may be used as a result of the path computation function.

5. Security Considerations

No additional security measure is required.

6. IANA Considerations

6.1. SR Capability Flag

IANA maintains a sub-registry, named "SR Capability Flag Field", within the "Path Computation Element Protocol (PCEP) Numbers" registry to manage the Flags field of the SR-PCE-CAPABILITY TLV. IANA is requested to make the following assignment:

Value	Description	Reference
TBD1	SID Algorithm Capability	This document

Table 1

6.2. SRv6 PCE Capability Flag

IANA was requested in [I-D.ietf-pce-segment-routing-ipv6] to create a sub-registry, named "SRv6 PCE Capability Flags", within the "Path Computation Element Protocol (PCEP) Numbers" registry to manage the Flags field of SRv6-PCE-CAPABILITY sub-TLV. IANA is requested to make the following assignment:

Value	Description	Reference
TBD2	SID Algorithm Capability	This document

Table 2

6.3. SR-ERO Flag

IANA maintains a sub-registry, named "SR-ERO Flag Field", within the "Path Computation Element Protocol (PCEP) Numbers" registry to manage the Flags field of the SR-ERO Subobject. IANA is requested to make the following assignment:

Value	Description	Reference
TBD3	SID Algorithm Flag	This document

Table 3

6.4. SRv6-ERO Flag

IANA was requested in [I-D.ietf-pce-segment-routing-ipv6], named "SRv6-ERO Flag Field", within the "Path Computation Element Protocol (PCEP) Numbers" registry to manage the Flags field of the SRv6-ERO subobject. IANA is requested to make the following assignment:

Value	Description	Reference
TBD4	SID Algorithm Flag	This document

Table 4

6.5. PCEP TLV Types

IANA is requested to allocate a new TLV type for the new LSPA TLV specified in this document.

Value	Description	Reference
TBD5	SID Algorithm	This document

Table 5

7. Normative 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", Work in Progress,
 Internet-Draft, draft-ietf-pce-segment-routing-ipv6-09, 27
 May 2021, <<https://www.ietf.org/internet-drafts/draft-ietf-pce-segment-routing-ipv6-09.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>>.
- [RFC5541] Le Roux, JL., Vasseur, JP., and Y. Lee, "Encoding of Objective Functions in the Path Computation Element Communication Protocol (PCEP)", RFC 5541, DOI 10.17487/RFC5541, June 2009, <<https://www.rfc-editor.org/info/rfc5541>>.
- [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>>.
- [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>>.

Appendix A. Contributors

Mike Koldychev
Cisco Systems
Kanata, Ontario
Canada

Email: mkoldych@cisco.com

Authors' Addresses

Alex Tokar
Cisco Systems, Inc.
Eurovea Central 3.
Pribinova 10
811 09 Bratislava
Slovakia

Email: atokar@cisco.com

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

Email: ssidor@cisco.com

Shaofu Peng
ZTE Corporation
No.50 Software Avenue
Nanjing
Jiangsu, 210012
China

Email: peng.shaofu@zte.com.cn

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

Email: msiva282@gmail.com

Tarek Saad
Juniper Networks

Email: tsaad@juniper.net

Shuping Peng
Huawei Technologies
Huawei Campus, No. 156 Beiqing Rd.
Beijing
100095
China

Email: pengshuping@huawei.com

Mahendra Singh Negi
RtBrick Inc
Bangalore
Karnataka
India

Email: mahend.ietf@gmail.com

PCE Working Group
Internet-Draft
Intended status: Standards Track
Expires: November 19, 2021

Y. Wang
A. Wang
China Telecom
May 18, 2021

PCEP Procedures and Extension for VLAN-based Traffic Forwarding
draft-wang-pce-vlan-based-traffic-forwarding-00

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

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 November 19, 2021.

Copyright Notice

Copyright (c) 2021 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<https://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of

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

Table of Contents

1. Introduction	2
2. Conventions used in this document	3
3. Terminology	3
4. Procedures for VLAN-based Traffic Forwarding	3
5. Capability Advertisement	4
6. PCEP message	4
6.1. The PCInitiate message	5
6.2. The PCRpt message	6
7. VXLAN-based traffic forwarding Procedures	7
7.1. Multiple BGP Session Establishment Procedures	7
7.2. BGP Prefix Advertisement Procedures	8
7.3. VLAN mapping info Advertisement Procedures	9
7.3.1. VLAN-Based forwarding info Advertisement Procedures	9
7.3.2. VLAN-Based crossing info Advertisement Procedures	10
8. New PCEP Objects	12
8.1. VLAN forwarding CCI Object	12
8.2. Peer IP Address TLVs	13
8.3. VLAN crossing CCI Object	14
9. Deployment Considerations	15
10. Security Considerations	15
11. IANA Considerations	15
11.1. Path Setup Type Registry	15
11.2. PCECC-CAPABILITY sub-TLV's Flag field	15
11.3. PCEP Object Types	15
11.4. PCEP-Error Object	15
12. Acknowledgement	16
13. Normative References	16
Authors' Addresses	17

1. Introduction

Based on the PCEP, a southbound interface protocol of the controller, the PCE can calculate the optimal path for various applications and sends it to the network equipment through the centralized path calculation mechanism, so as to control the packet forwarding and make the separation of path calculation and establishment function.

With the large scale deployment of Ethernet interface, it is possible to use the info contained in the Layer2 message to simplify the processing of a distributed control plane. 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. 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. Compared with other traffic assurance technologies such as mpls or srv6, the VLAN-based traffic forwarding mechanism uses a completely new address space which will not conflict with other existing protocols. 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: Label Switching Path
- o PST: Path Setup Type

4. Procedures for VLAN-based Traffic Forwarding

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 relabeled with new VLAN tag and then be forwarded.

5. 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:

- 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 [I-D.ietf-pce-extension-for-pce-controller]

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 .

[I-D.ietf-pce-pcep-extension-for-pce-controller] 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 enxtends 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[I-D.ietf-pce-pcep-extension-for-pce-controller] 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
[I-D.ietf-pce-pcep-extension-for-pce-controller].
<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 [I-D.ietf-pce-pcep-extension-for-pce-controller]. 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 acknowledged 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 [I-D.ietf-pce-pcep-extension-for-pce-controller]. 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. VXLAN-based traffic forwarding Procedures

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 informations 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:

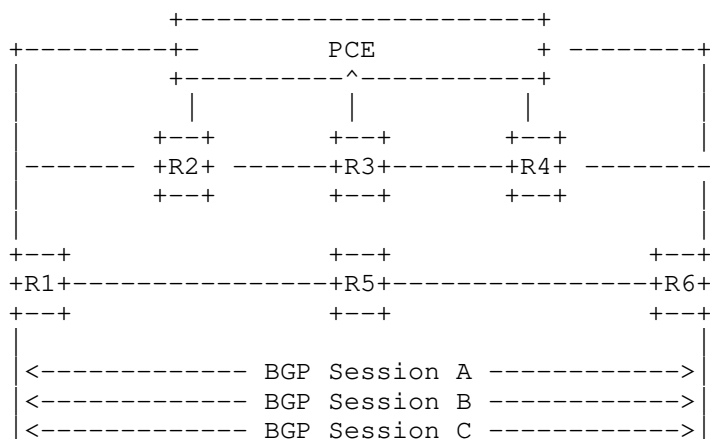


Figure 1: BGP Session Establishment Procedures

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. The specific process is as follows,:

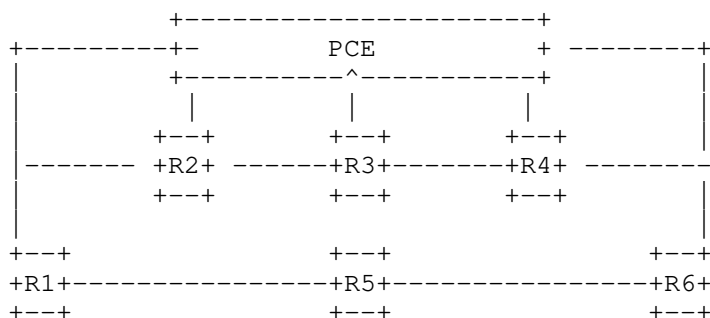


Figure 2: BGP Prefix Advertisement Procedures

Through BGP protocol, the ingress device can learn different BGP prefix of the egress device based on the different BGP sessions.

7.3. VLAN mapping info Advertisement Procedures

After the BGP prefix for different BGP session are successfully advertised, informations 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 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 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 the packet is tagged and successfully sent, the PCC should report the result via the PCRpt messages, with VLAN forwarding CCI Object and the corresponding SRP object included.

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.

When PCC withdraws the VLAN-Based forwarding info that indicated by this object successfully, it should report the result via the PCRpt message, with the corresponding SRP and CCI object included.

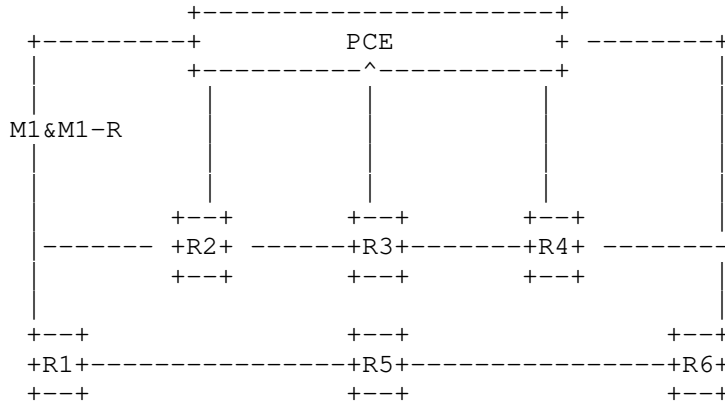


Figure 3: VLAN-Based forwarding info Advertisement Procedures for Ingress PCC

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	PCE/R1	PCInitiate	CC-ID=X1
M1-R		PCRpt	VLAN Forwarding CCI Object (Peer_IP=R6_A, 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 is shown below, using PCInitiate and PCRpt message pair.

After the process of VLAN-Based forwarding info advertisement mentioned above, the PCC will form a VLAN-crossing routing table 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 will be replaced by a new out-VLAN tag 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 the packet is tagged and successfully sent to the specific subinterface, the PCC should report the result via the PCRpt messages, with the corresponding SRP and CCI object included.

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.

When PCC withdraws the VLAN-Based crossing info that indicated by this object successfully, it should report the result via the PCRpt message, 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.

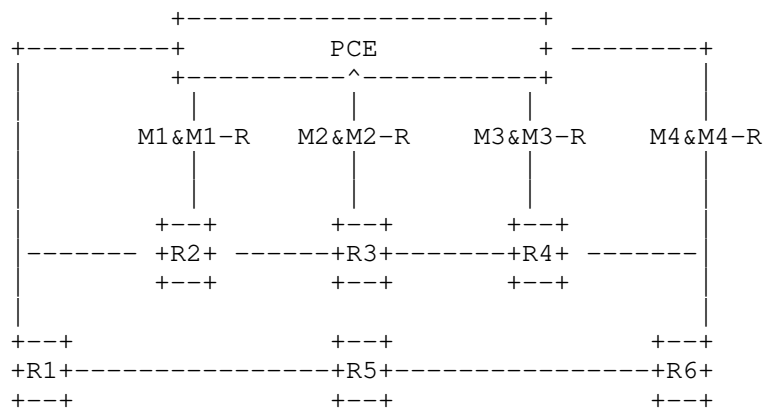


Figure 4: VLAN-Based crossing info Advertisement Procedures for transit PCC and egress PCC

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 VLAN crossing CCI Object (IN_VLAN_ID=VLAN_R1_R2,OUT_VLAN_ID=VLAN_R2_R3)
M2 M2-R	PCE/R3	PCInitiate PCRpt	CC-ID=X1 VLAN crossing CCI Object (IN_VLAN_ID=VLAN_R2_R3,OUT_VLAN_ID=VLAN_R3_R4)
M3 M3-R	PCE/R4	PCInitiate PCRpt	CC-ID=X1 VLAN crossing CCI Object (IN_VLAN_ID=VLAN_R3_R4,OUT_VLAN_ID=VLAN_R4_R6)
M4 M4-R	PCE/R6	PCInitiate PCRpt	CC-ID=X1 VLAN crossing CCI Object (IN_VLAN_ID=VLAN_R4_R6,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 [I-D.ietf-pce-pcep-extension-for-pce-controller]. 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].

8.1. 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.

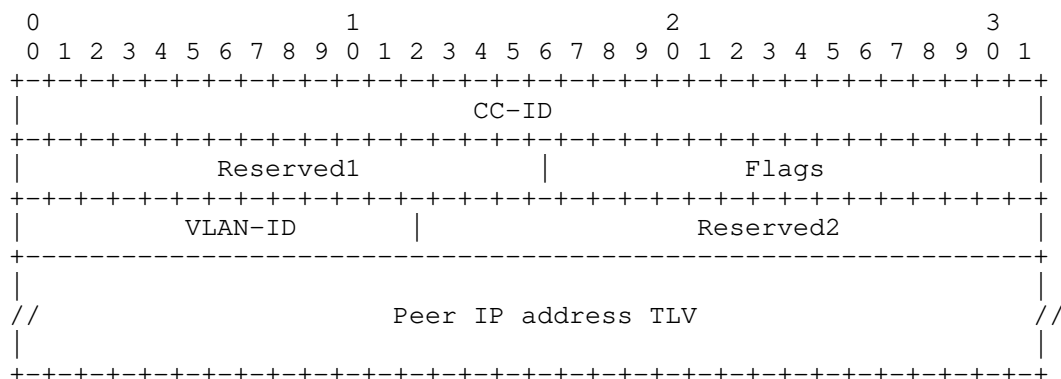


Figure 5: VLAN forwarding CCI Object

The fields in the CCI object are as follows:

CC-ID: is as described in [I-D.ietf-pce-pcep-extension-for-pce-controller]. 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.

The Peer IP Address TLV[RFC8231]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. Peer IP 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 PCError 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.

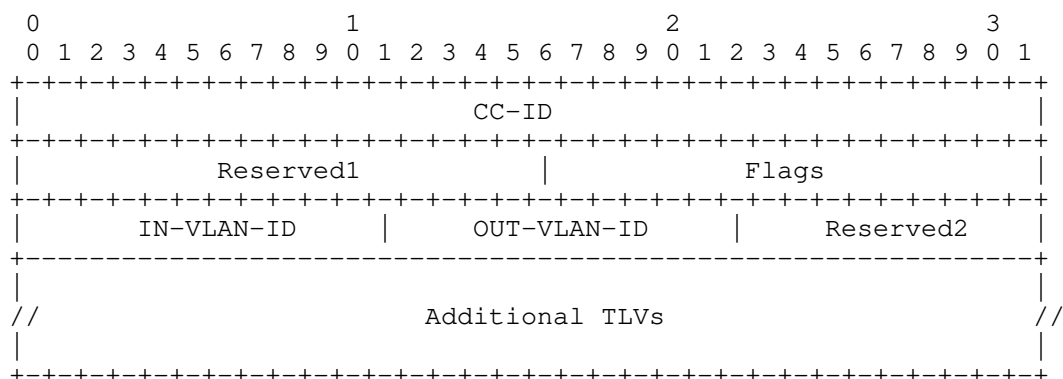


Figure 6: VLAN Crossing CCI Object

CC-ID: is as described in [I-D.ietf-pce-pcep-extension-for-pce-controller]. 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 no flag bits are defined.

IN-VLAN ID (12 bits): The ID of the VLAN forwarding path which is used to identify the traffic that needs to be protected.

OUT-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 labeled 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.

9. Deployment Considerations

10. Security Considerations

11. IANA Considerations

11.1. 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

11.2. PCECC-CAPABILITY sub-TLV's Flag field

[I-D.ietf-pce-pcep-extension-for-pce-controller] 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

11.3. PCEP Object Types

IANA is requested to allocate new registry for the PCEP Object Type:

Object-Class Value	Name	Reference
44	CCI Object-Type	This document
	TBD8: VLAN forwarding CCI	
	TBD10: VLAN crossing CCI	

11.4. 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 TBD9: Invalid egress PCC information	This document This document

12. Acknowledgement

13. Normative References

- [I-D.ietf-pce-pcep-extension-for-pce-controller]
Li, Z., Peng, S., Negi, M. S., Zhao, Q., and C. Zhou,
"PCEP Procedures and Protocol Extensions for Using 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-13 (work in progress), March
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>>.

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

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

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

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 4, 2022.

Copyright Notice

Copyright (c) 2022 IETF Trust and the persons identified as the document authors. All rights reserved.

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

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

Table of Contents

1. Introduction	2
2. Conventions used in this document	3
3. Terminology	4
4. Procedures for VLAN-based Traffic Forwarding	4
5. Capability Advertisement	5
6. PCEP message	5
6.1. The PCInitiate message	6
6.2. The PCRpt message	7
7. VSP Operations	8
8. VXLAN-based traffic forwarding Procedures	12
8.1. Multiple BGP Session Establishment Procedures	12
8.2. BGP Prefix Advertisement Procedures	13
8.3. VLAN mapping info Advertisement Procedures	13
8.3.1. VLAN-Based forwarding info Advertisement Procedures	13
8.3.2. VLAN-Based crossing info Advertisement Procedures	15
9. New PCEP Objects	17
9.1. VLAN forwarding CCI Object	17
9.2. Address TLVs	19
9.3. VLAN crossing CCI Object	19
10. Deployment Considerations	20
11. Security Considerations	20
12. IANA Considerations	20
12.1. Path Setup Type Registry	20
12.2. PCECC-CAPABILITY sub-TLV's Flag field	21
12.3. PCEP Object Types	21
12.4. PCEP-Error Object	21
13. Acknowledgement	21
14. Normative References	22
Authors' Addresses	23

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 achieve 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: Label 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 relabeled 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 level1, 0180:C200:0015 for ISIS level2) 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 traffic forwarding 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.

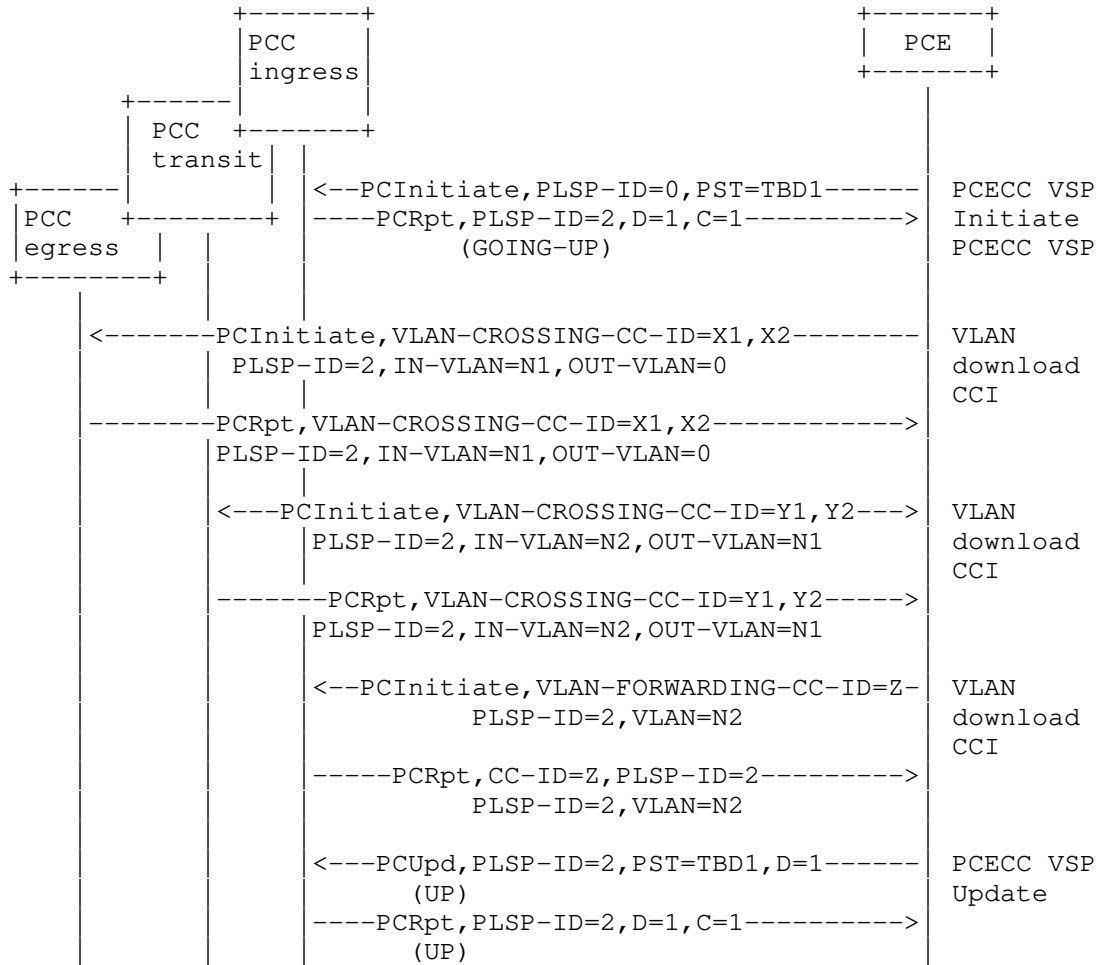
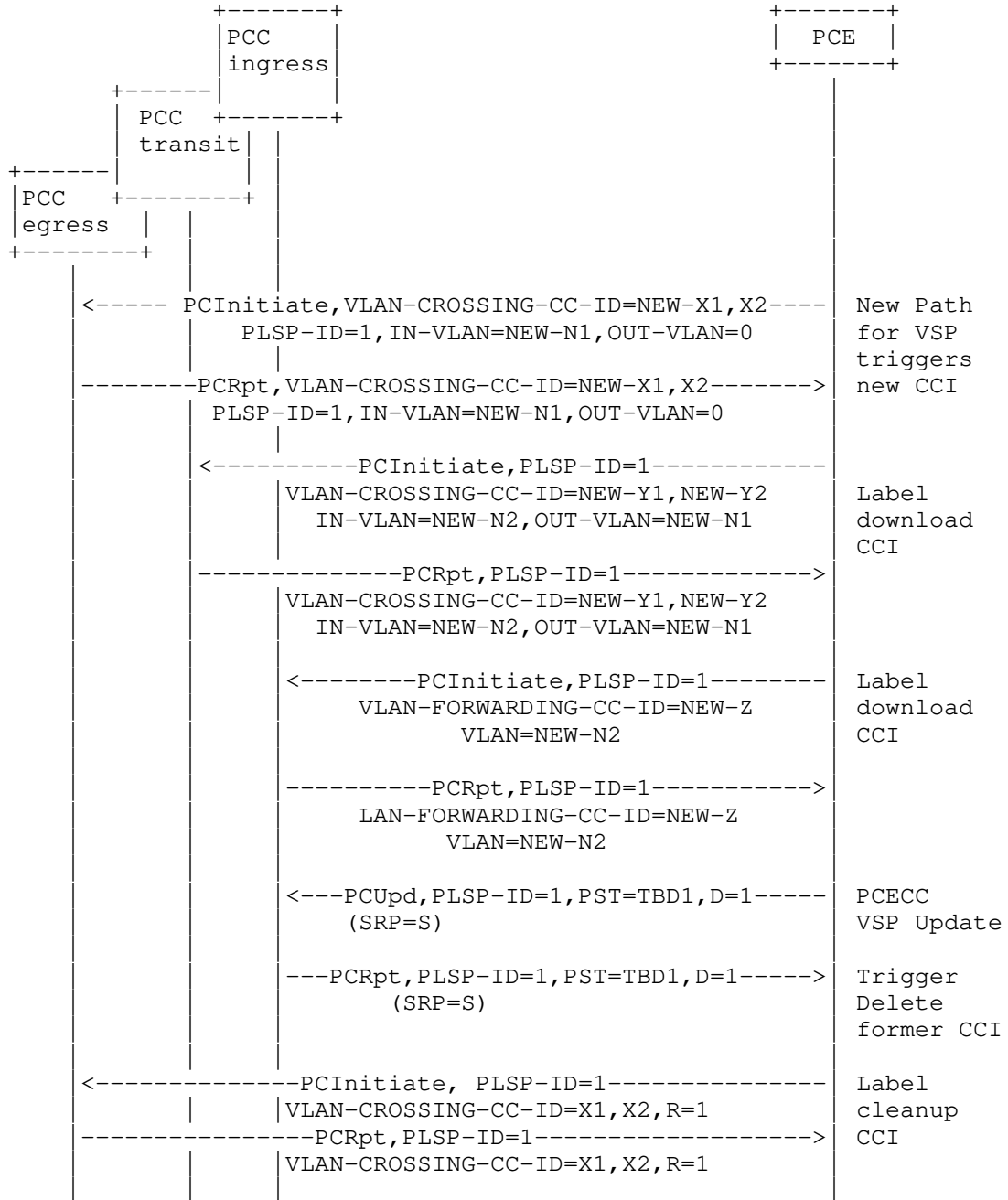


Figure 1: PCE-Initiated PCECC VSP

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.



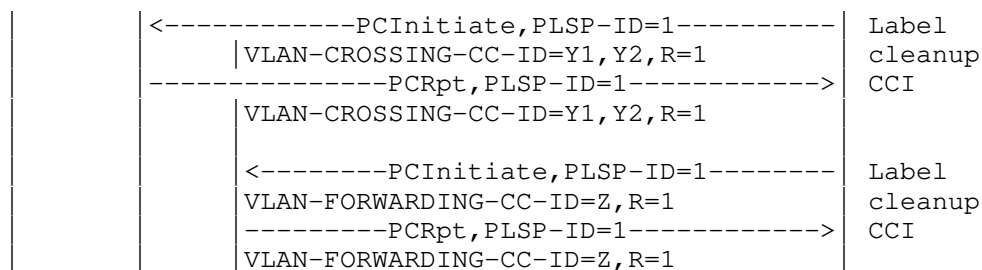


Figure 2: PCECC VSP Update

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:

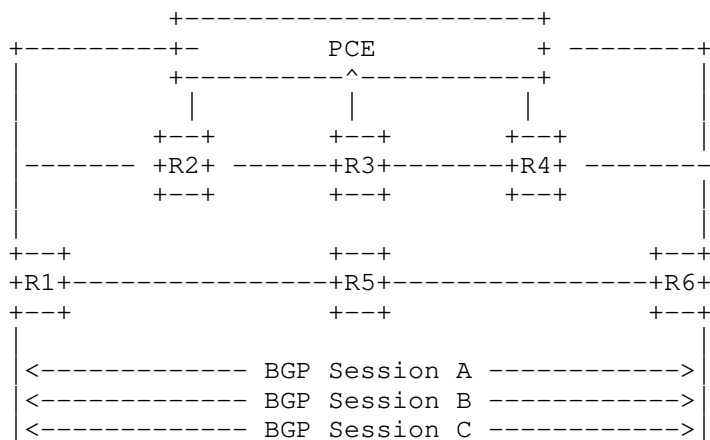


Figure 3: BGP Session Establishment Procedures

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.

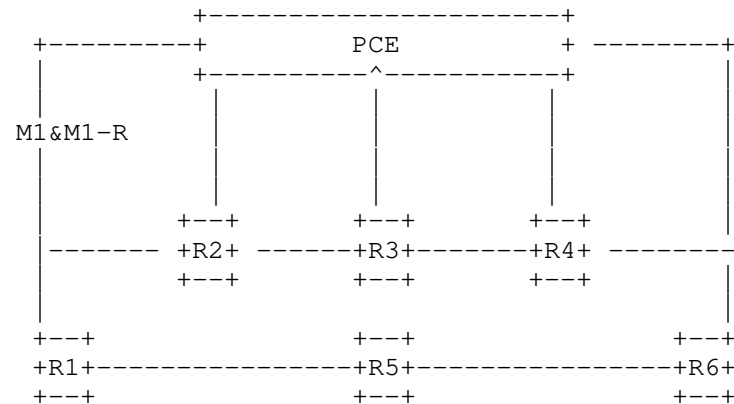


Figure 4: VLAN-Based forwarding info Advertisement Procedures for Ingress PCC

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.

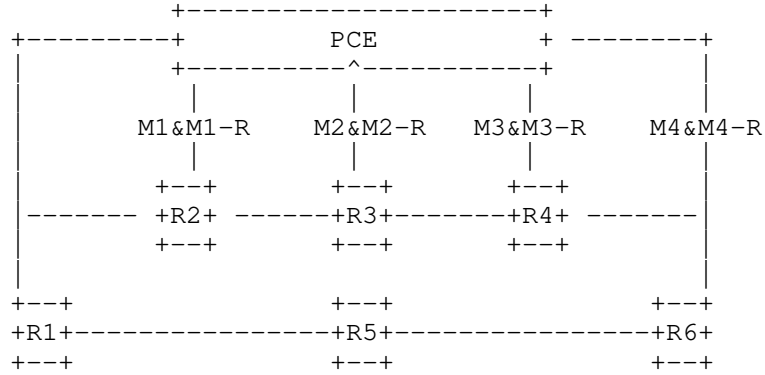


Figure 5: VLAN-Based crossing info Advertisement Procedures
for transit PCC and egress PCC

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.

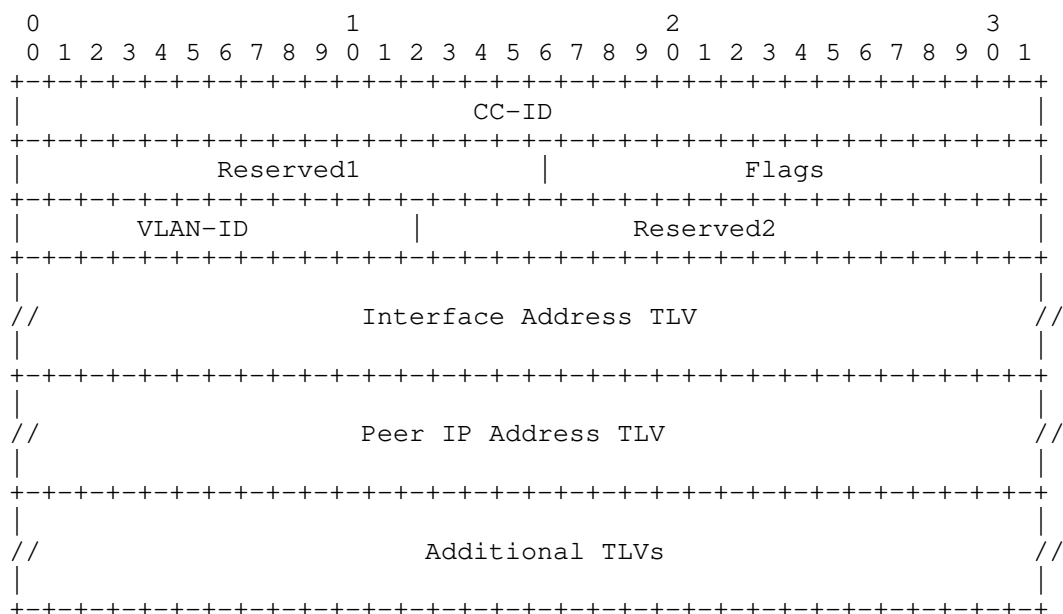


Figure 6: VLAN Forwarding CCI Object

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.

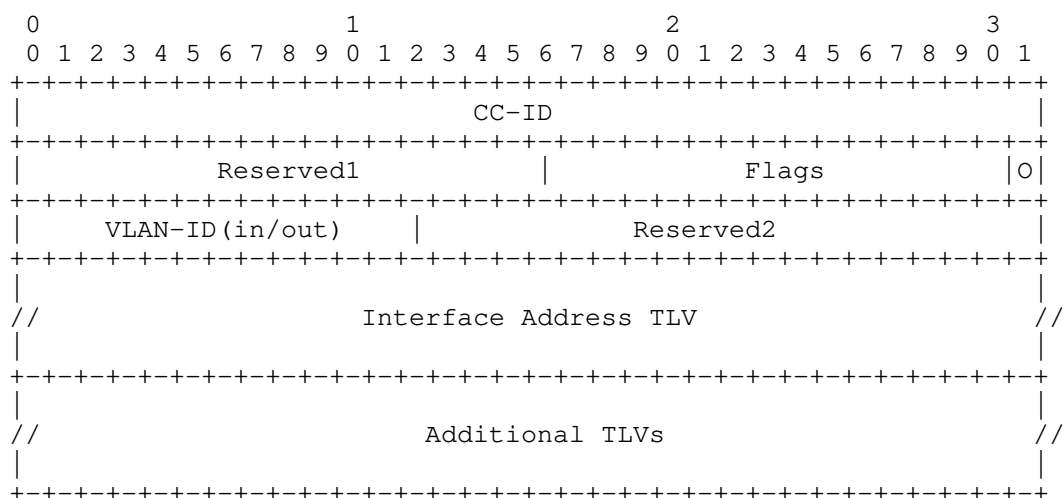


Figure 7: VLAN Crossing CCI Object

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 labeled 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	This document
	TBD8: VLAN forwarding CCI	
	TBD10: VLAN crossing CCI	

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