

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.

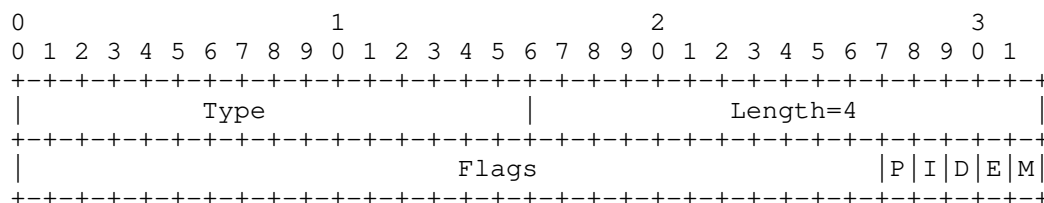


Fig. 1 IFIT-CAPABILITY TLV Format

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:

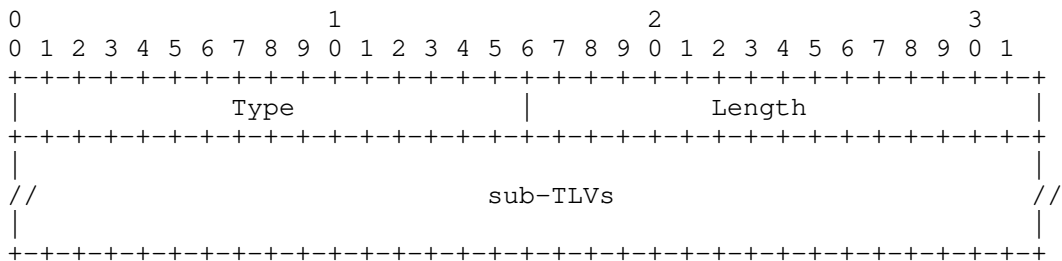


Fig. 2 IFIT-ATTRIBUTES TLV Format

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:

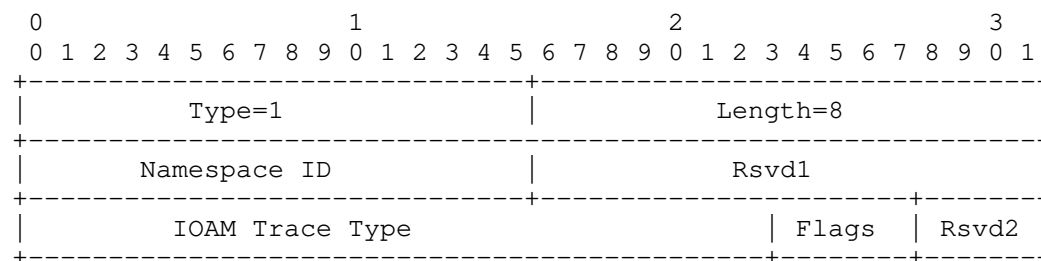


Fig. 4 IOAM Pre-allocated Trace Option Sub-TLV

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:

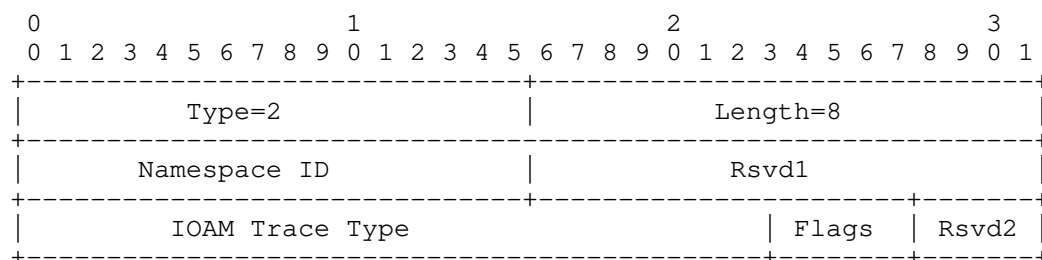


Fig. 5 IOAM Incremental Trace Option Sub-TLV

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:

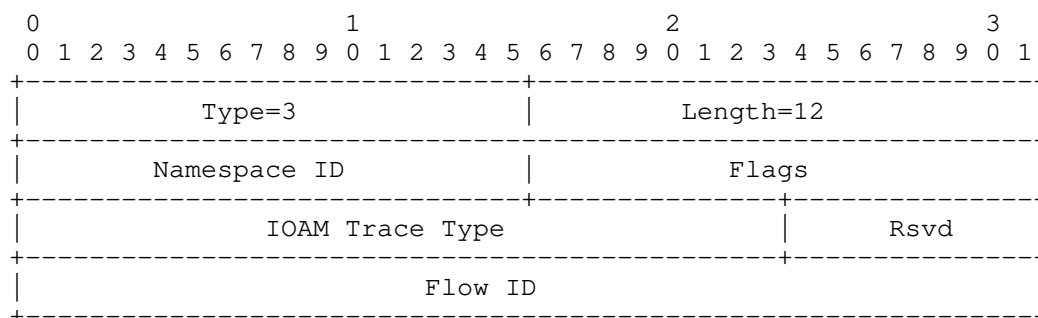


Fig. 6 IOAM Directly Export Option Sub-TLV

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:

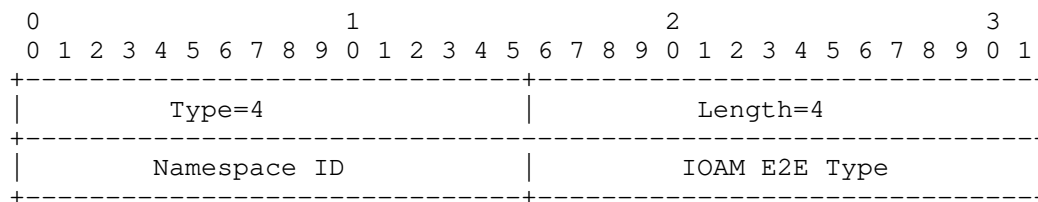


Fig. 7 IOAM Edge-to-Edge Option Sub-TLV

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:

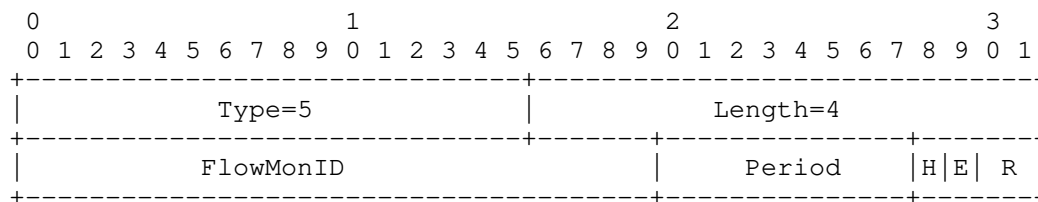


Fig. 8 Enhanced Alternate Marking Sub-TLV

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.

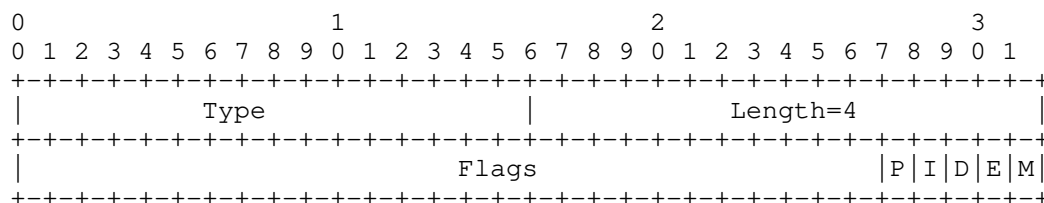


Fig. 1 IFIT-CAPABILITY TLV Format

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:

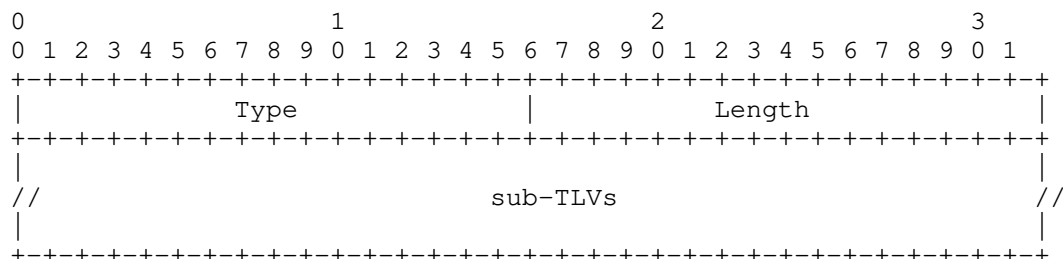


Fig. 2 IFIT-ATTRIBUTES TLV Format

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:

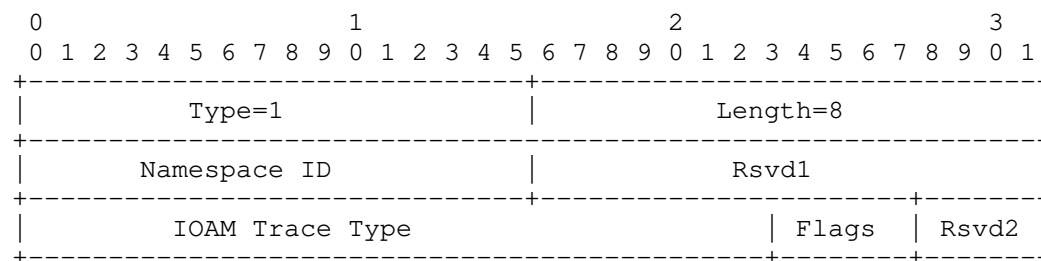


Fig. 4 IOAM Pre-allocated Trace Option Sub-TLV

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:

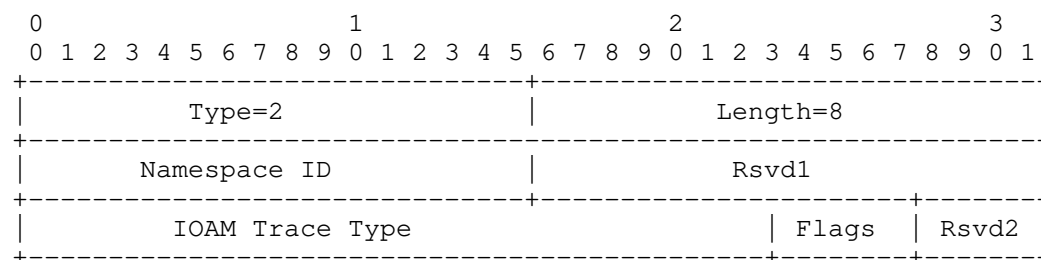


Fig. 5 IOAM Incremental Trace Option Sub-TLV

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:

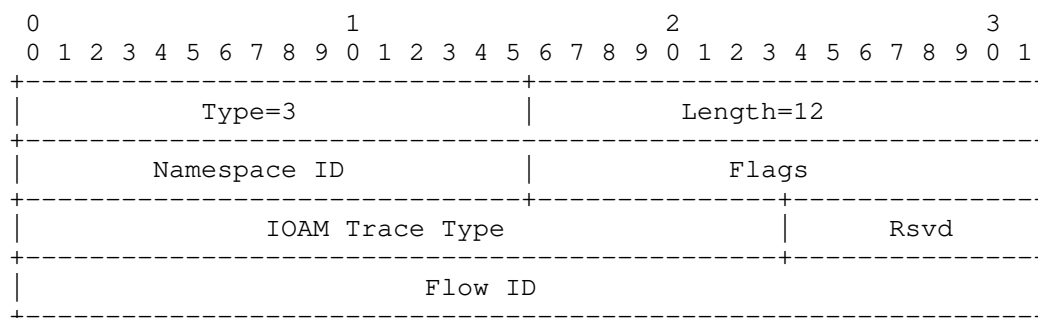


Fig. 6 IOAM Directly Export Option Sub-TLV

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:

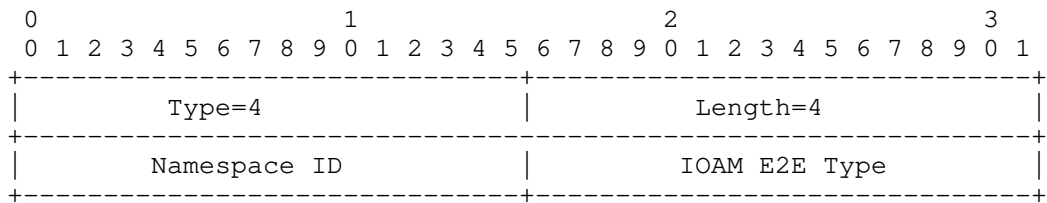


Fig. 7 IOAM Edge-to-Edge Option Sub-TLV

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-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:

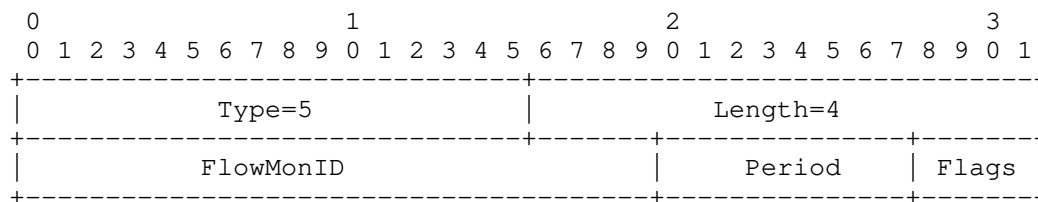


Fig. 8 Enhanced Alternate Marking Sub-TLV

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.

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: 25 April 2022

H. Chen
M. McBride
Futurewei
M. Toy
G. Mishra
Verizon Inc.
A. Wang
China Telecom
Z. Li
Y. Liu
China Mobile
B. Khasanov
Yandex LLC
L. Liu
Fujitsu
X. Liu
Volta Networks
22 October 2021

Path Ingress Protections
draft-chen-pce-sr-ingress-protection-06

Abstract

This document describes extensions to Path Computation Element (PCE) communication Protocol (PCEP) for fast protecting the ingress nodes of two types of paths or tunnels, which are Segment Routing (SR) paths and Bit Index Explicit Replication Tree/Traffic Engineering (BIER-TE) paths. The extensions comprise a foundation for protecting the ingress nodes of different types of paths. Based on this, the ingress protection of a new type of paths can be easily supported.

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 25 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
1.1. Terminologies	3
2. Path Ingress Protection Examples	4
2.1. SR Path Ingress Protection Example	4
2.2. BIER-TE Path Ingress Protection Example	5
3. Behavior around Ingress Failure	6
3.1. Source Detect	6
3.2. Backup Ingress Detect	6
3.3. Both Detect	7
4. Extensions to PCEP	7
4.1. Capabilities for Ingress Protection	7
4.1.1. Capability for Ingress Protection with Backup Ingress	7
4.1.2. Capability for Ingress Protection with Traffic Source	9
4.2. Extensions for Backup Ingress and Traffic Source	10
4.2.1. Extensions for Backup Ingress	10
4.2.2. Extensions for Traffic Source	16
5. Security Considerations	19
6. Acknowledgements	19
7. IANA Considerations	19
8. References	19
8.1. Normative References	19
8.2. Informative References	19
Authors' Addresses	20

1. Introduction

The fast protection of a transit node in each type of paths or tunnels have been proposed. For example, the fast protection of a transit node in a Segment Routing (SR) path or tunnel is described in [I-D.ietf-rtgwg-segment-routing-ti-lfa]. 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 an SR path/tunnel, a BIER-TE path/tunnel, or other type of paths/tunnels.

This document fills that void and specifies protocol extensions to Path Computation Element (PCE) communication Protocol (PCEP) [RFC5440] and [RFC9050] for fast protecting the ingress nodes of two types of paths: SR paths and BIER-TE paths. The extensions comprise a foundation for protecting the ingress nodes of different types of paths. Based on this, the ingress protection of a new type of paths can be easily supported.

Ingress node and ingress, fast protection and protection, path ingress protection and ingress protection, SR path and SR tunnel, as well as BIER-TE path and BIER-TE tunnel will be used exchangeably in the following sections.

1.1. Terminologies

The following terminologies are used in this document.

PCE: Path Computation Element or Path Computation Element server

PCEP: PCE communication Protocol

PCC: Path Computation Client

BIER: Bit Index Explicit Replication

BIFT: Bit Index Forwarding Table

CE: Customer Edge

PE: Provider Edge

TE: Traffic Engineering

SR: Segment Routing

LFA: Loop-Free Alternate

TI-LFA: Topology Independent LFA

BFD: Bidirectional Forwarding Detection

VPN: Virtual Private Network

L3VPN: Layer 3 VPN

FIB: Forwarding Information Base

2. Path Ingress Protection Examples

This section shows two examples of path ingress protection. One is SR path ingress protection, and the other is BIER-TE path ingress protection.

2.1. SR Path Ingress Protection Example

Figure 1 shows an example of protecting ingress PE1 of a SR path, which is from ingress PE1 to egress PE3 and represented by *** in the figure.

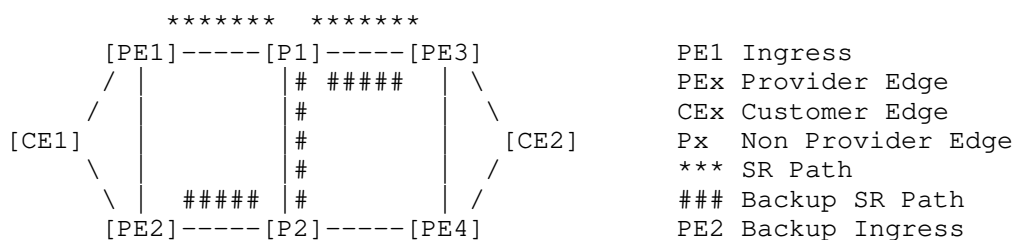


Figure 1: Protecting Ingress PE1 of SR Path

In normal operations, CE1 sends the traffic with destination PE3 to ingress PE1, which imports the traffic into the SR path.

When CE1 detects the failure of ingress PE1, it switches the traffic to backup ingress PE2, which imports the traffic from CE1 into a backup SR path. The backup path is from the backup ingress PE2 to the egress PE3 and represented by ### in the figure. When the traffic is imported into the backup path, it is sent to the egress PE3 along the path.

When the PCC of the traffic source receives the information about the backup ingress, the primary ingress and the traffic, 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.

3. 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.

3.1. Source Detect

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

3.2. Backup Ingress Detect

The traffic source (e.g., CE1) always sends the traffic to both the ingress (e.g., PE1) of the primary path such as 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 path such as the backup BIER-TE path in normal operations. When it detects the failure of the ingress of the primary path, it imports the traffic from the source into the backup 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 path, it SHOULD consider this.

3.3. 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 path such as 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 path such as the backup BIER-TE path in normal operations. When it detects the failure of the ingress of the primary path, it imports the traffic from the source into the backup path.

4. Extensions to PCEP

A PCC runs on each of the edge nodes such as PEs of a network normally. A PCE runs on a server as a controller to communicate with PCCs. PCE and PCCs work together to support protection for the ingress of a path. The path is a SR path, a BIER-TE path, or a path of another type.

4.1. Capabilities for Ingress Protection

4.1.1. 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 each of different types of paths.

A new sub-TLV called INGRESS_PROTECTION_CAPABILITY is defined. It is included in the PATH_SETUP_TYPE_CAPABILITY TLV with PST = TBD1 (suggested value 2 for path ingress protection) 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.

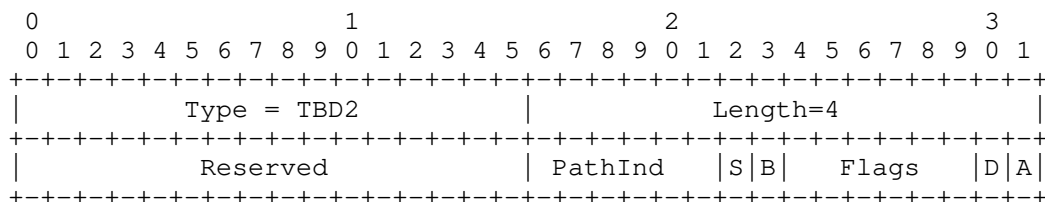


Figure 3: INGRESS_PROTECTION_CAPABILITY sub-TLV

Type: TBD2 is to be assigned by IANA.

Length: 4.

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

PathInd: 1 octet. Indicators for the types of paths whose ingress protections are supported. Two indicators are defined.

- o S : S = 1 indicating that the ingress protection of a SR path is supported.
- o B : B = 1 indicating that the ingress protection of a BIER-TE path is supported.

Flags: 1 octet. Two flags are defined.

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

A PCC, which supports ingress protection for different types of paths, sends a PCE an Open message containing INGRESS_PROTECTION_CAPABILITY sub-TLV. This sub-TLV indicates that the PCC is capable of supporting the ingress protection for the types of paths.

For example, if a PCC supports ingress protection for SR path and BIER-TE path, the PCC sends a PCE an Open message containing INGRESS_PROTECTION_CAPABILITY sub-TLV with S = 1 and B = 1.

A PCE, which supports ingress protection for different types of paths, sends a PCC an Open message containing INGRESS_PROTECTION_CAPABILITY sub-TLV. This sub-TLV indicates that the PCE is capable of supporting the ingress protection for the types of paths.

If both a PCC and a PCE support 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 an INGRESS_PROTECTION_CAPABILITY sub-TLV.

If a PCE receives an Open message from a PCC without a INGRESS_PROTECTION_CAPABILITY sub-TLV indicating PCC's support for the ingress protection of a type of paths, then the PCE MUST not send the PCC any request for ingress protection of the type of paths.

If a PCC receives an Open message from a PCE without a INGRESS_PROTECTION_CAPABILITY sub-TLV indicating PCE's support for the ingress protection of a type of paths, then the PCC MUST ignore any request for ingress protection of the type of paths 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 path, the PCC MUST let the forwarding entry for the backup path be Active.

4.1.2. 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 ingress protection.

The PCECC-CAPABILITY sub-TLV defined in [RFC9050] 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.

4.2. Extensions for Backup Ingress and Traffic Source

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 path, or

B2: import the traffic with possible service ID into the backup 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.

4.2.1. Extensions for Backup Ingress

For the packets from the traffic source, if the primary ingress (i.e., the ingress of the primary path) encapsulates the packets with a service ID or label into the path, the backup ingress MUST have this service ID or label and encapsulates the packets with the service ID or label into the backup 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 sub-TLV called INGRESS_PROTECTION is defined. When a PCE sends a PCC a PCInitiate message for initiating a backup path to protect the primary ingress node of a primary path, the message contains this TLV in the RP/SRP object. Its format is illustrated below.

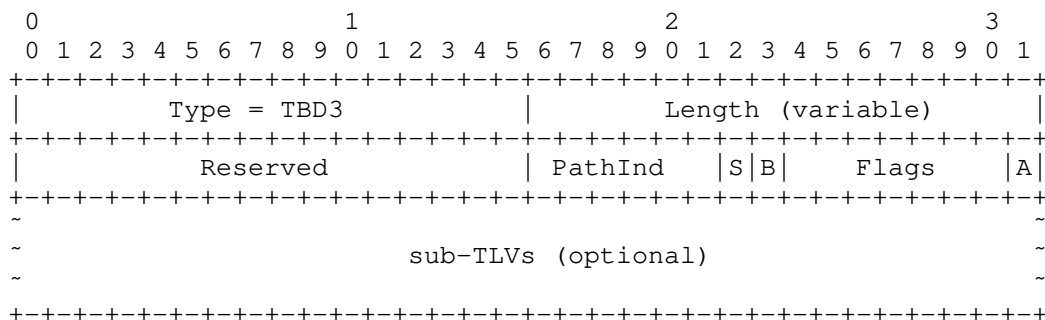


Figure 4: INGRESS_PROTECTION sub-TLV

Type: TBD3 is to be assigned by IANA.

Length: Variable.

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

PathInd: 1 octet. Indicating for the types of paths whose ingress nodes are protected.

- o S : S = 1 indicating the ingress protection of a SR path.
- o B : B = 1 indicating the ingress protection of a BIER-TE path.

Flags: 1 octet. One flag 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 path 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 path 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.

Three optional sub-TLVs are defined: Primary-Ingress sub-TLV, Service sub-TLV, and Traffic-Description sub-TLV. The Traffic-Description sub-TLV describes the traffic to be imported into the backup SR path. 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.

4.2.1.1. Primary-Ingress sub-TLV

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

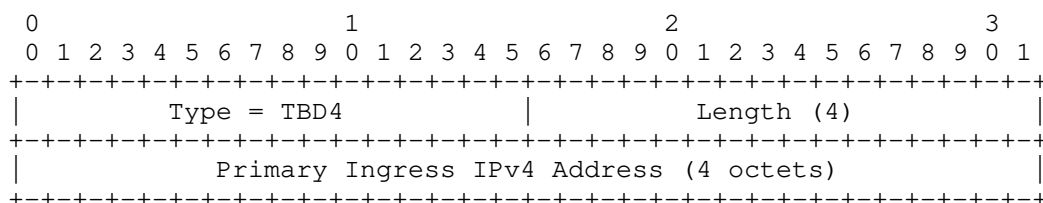


Figure 5: Primary Ingress IPv4 Address sub-TLV

Type: TBD4 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 path.

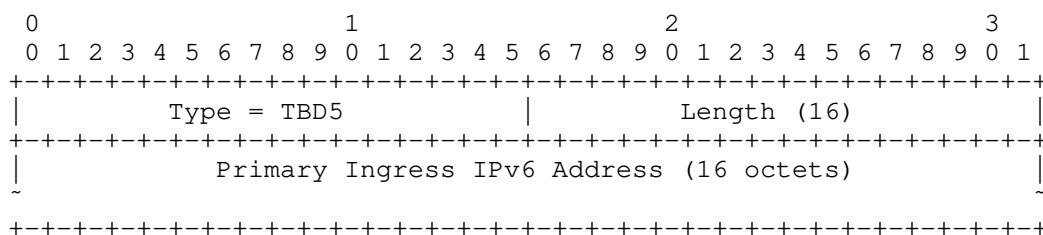


Figure 6: Primary Ingress IPv6 Address sub-TLV

Type: TBD5 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 path.

4.2.1.2. Service sub-TLV

A Service sub-TLV contains a service ID or label to be added into a packet to be carried by a path. 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.

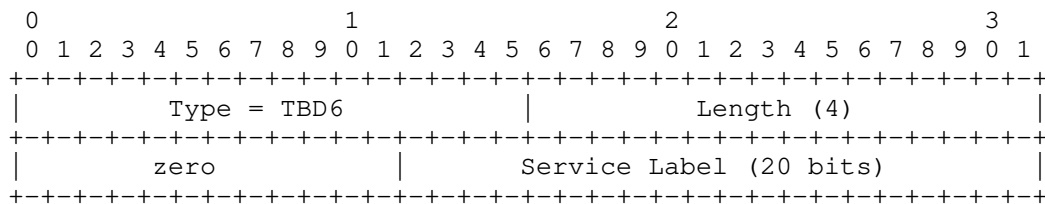


Figure 7: Service Label sub-TLV

Type: TBD6 is to be assigned by IANA.

Length: 4.

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

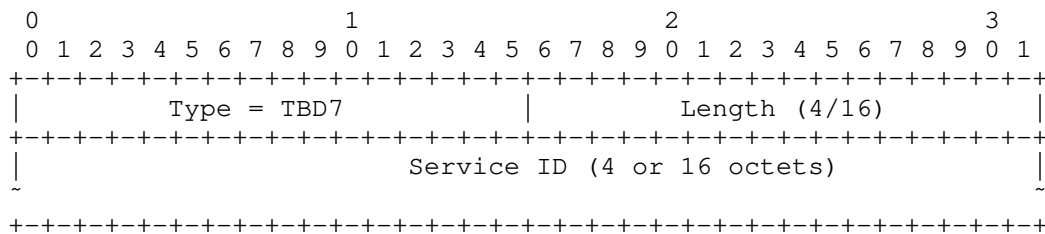


Figure 8: Service ID sub-TLV

Type: TBD7 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.

4.2.1.3. Traffic-Description sub-TLV

A Traffic-Description sub-TLV describes the traffic to be imported into a backup SR path. Its format is illustrated below.

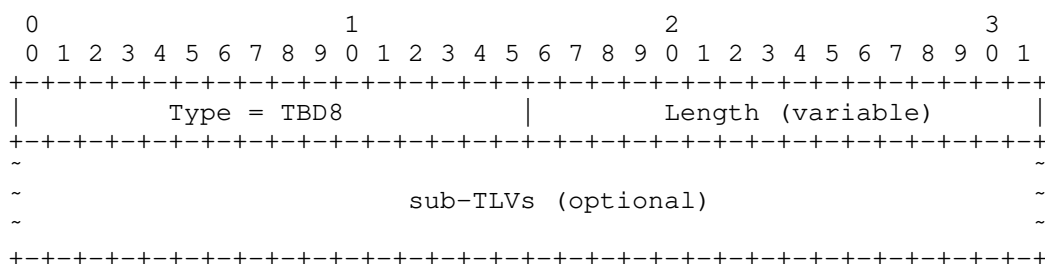


Figure 9: Traffic-Description sub-TLV

Type: TBD8 is to be assigned by IANA.

Length: Variable.

Two optional sub-TLVs are defined. One is FEC sub-TLV and the other interface sub-TLV.

A FEC sub-TLV describes the traffic to be imported into the backup path. It is an IP prefix with an optional virtual network ID. It has two formats: one for IPv4 and the other for IPv6, which are illustrated below.

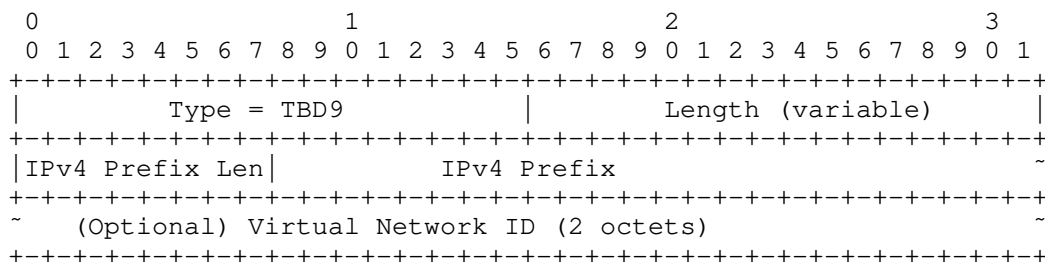


Figure 10: IPv4 FEC sub-TLV

Type: TBD9 is to be assigned by IANA.

Length: Variable.

IPv4 Prefix Len: Indicates the length of the IPv4 Prefix.

IPv4 Prefix: IPv4 Prefix rounded to octets.

Virtual Network ID: 2 octets. This is optional. It indicates the ID of a virtual network.

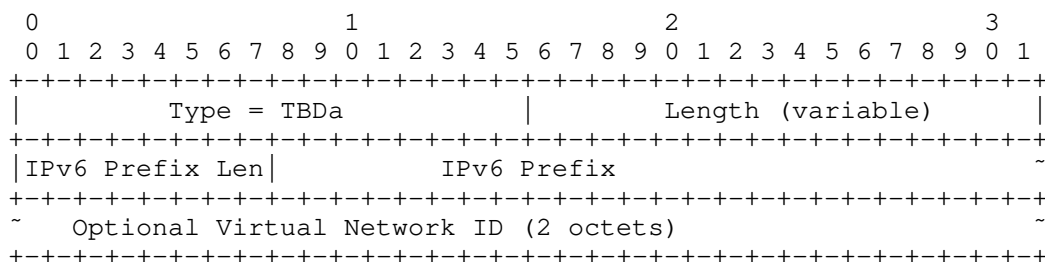


Figure 11: IPv6 FEC sub-TLV

Type: TBDA is to be assigned by IANA.

Length: Variable.

IPv6 Prefix Len: Indicates the length of the IPv6 Prefix.

IPv6 Prefix: IPv6 Prefix rounded to octets.

Virtual Network ID: 2 octets. This is optional. It indicates the ID of a virtual network.

An Interface sub-TLV indicates the interface from which the traffic is received and imported into the backup path. It has three formats: one for interface index, the other two for IPv4 and IPv6 address, which are illustrated below.

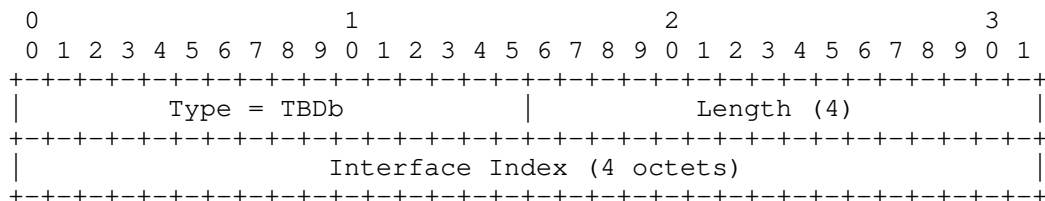


Figure 12: Interface Index sub-TLV

Type: TBDb is to be assigned by IANA.

Length: 4.

Interface Index: 4 octets. It indicates the index of an interface.

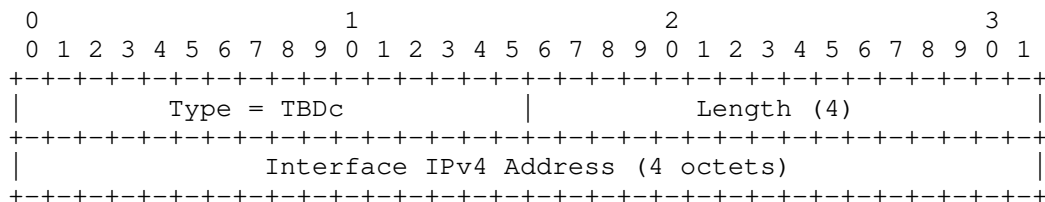


Figure 13: Interface IPv4 Address sub-TLV

Type: TBDc is to be assigned by IANA.

Length: 4.

Interface IPv4 Address: 4 octets. It represents the IPv4 address of an interface.

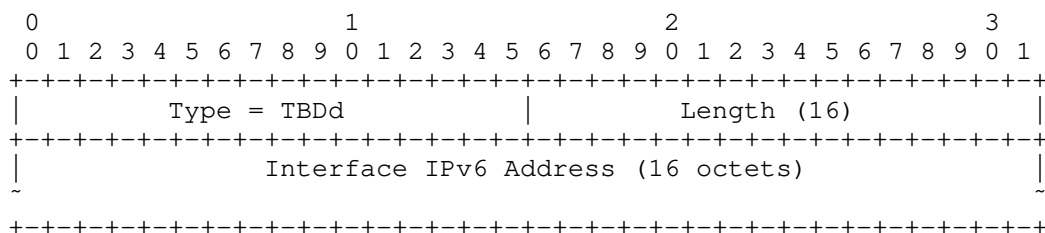


Figure 14: Interface IPv6 Address sub-TLV

Type: TBDd is to be assigned by IANA.

Length: 16.

Interface IPv6 Address: 16 octets. It represents the IPv6 address of an interface.

4.2.2. 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 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 [RFC9050] for a PCE as a controller to send instructions for LSPs to a PCC. This document defines a new object-type (TBDt) for 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.

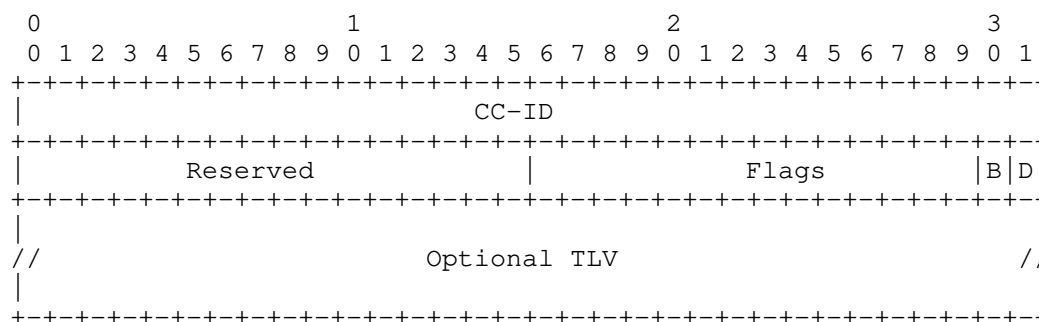


Figure 15: INGRESS-PROTECTION Object Body

CC-ID: It is the same as described in [RFC9050].

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, Traffic-Description TLV 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 Traffic-Description sub-TLV defined above is used as a TLV to contain the information about the traffic for a SR path 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 for a BIER-TE path in the object. A new TLV, called backup ingress TLV, is defined to contain the information about the backup ingress in the object.

4.2.2.1. Backup-Ingress TLV

A Backup-Ingress TLV indicates the IP address of the ingress node of a backup path. 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.

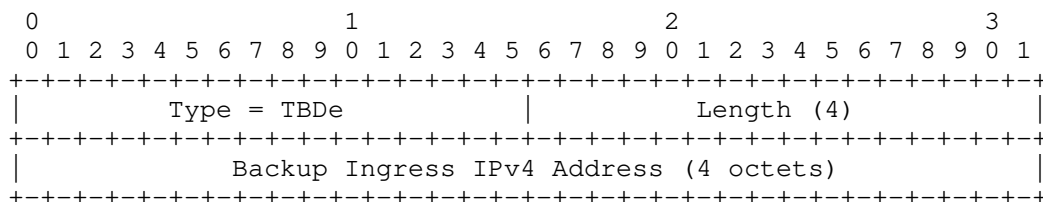


Figure 16: Backup Ingress IPv4 Address TLV

Type: TBDe is to be assigned by IANA.

Length: 4.

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

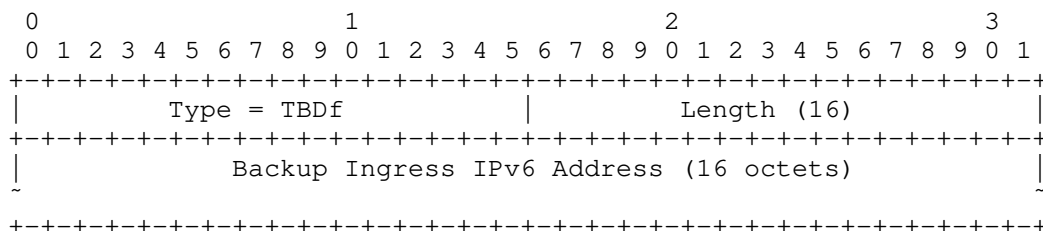


Figure 17: Backup Ingress IPv6 Address TLV

Type: TBDf 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.

5. Security Considerations

TBD

6. Acknowledgements

The authors of this document would like to thank Dhruv Dhody and Robin Li for their reviews and comments.

7. IANA Considerations

TBD

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>>.
- [RFC7356] Ginsberg, L., Previdi, S., and Y. Yang, "IS-IS Flooding Scope Link State PDUs (LSPs)", RFC 7356, DOI 10.17487/RFC7356, September 2014, <<https://www.rfc-editor.org/info/rfc7356>>.
- [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>>.
- [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>>.

8.2. 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-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>>.

[I-D.ietf-rtgwg-segment-routing-ti-lfa]

Litkowski, S., Bashandy, A., Filsfils, C., Francois, P., Decraene, B., and D. Voyer, "Topology Independent Fast Reroute using Segment Routing", Work in Progress, Internet-Draft, draft-ietf-rtgwg-segment-routing-ti-lfa-07, 29 June 2021, <<https://www.ietf.org/archive/id/draft-ietf-rtgwg-segment-routing-ti-lfa-07.txt>>.

[RFC5462] Andersson, L. and R. Asati, "Multiprotocol Label Switching (MPLS) Label Stack Entry: "EXP" Field Renamed to "Traffic Class" Field", RFC 5462, DOI 10.17487/RFC5462, February 2009, <<https://www.rfc-editor.org/info/rfc5462>>.

Authors' Addresses

Huaimo Chen
Futurewei
Boston, MA,
United States of America

Email: Huaimo.chen@futurewei.com

Mike McBride
Futurewei

Email: michael.mcbride@futurewei.com

Mehmet Toy
Verizon Inc.
United States of America

Email: mehmet.toy@verizon.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

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

Email: wangaj3@chinatelecom.cn

Zhenqiang Li
China Mobile
32 Xuanwumen West Ave, Xicheng District
Beijing
100053
China

Email: lizhengqiang@chinamobile.com

Yisong Liu
China Mobile

Email: liuyisong@chinamobile.com

Boris Khasanov
Yandex LLC
Moscow

Email: bhassanov@yahoo.com

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

Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: 17 May 2022

H. Chen
M. McBride
Futurewei
M. Toy
G. Mishra
Verizon Inc.
A. Wang
China Telecom
Z. Li
Y. Liu
China Mobile
B. Khasanov
Yandex LLC
L. Liu
Fujitsu
X. Liu
Volta Networks
13 November 2021

Path Ingress Protections
draft-chen-pce-sr-ingress-protection-07

Abstract

This document describes extensions to Path Computation Element (PCE) communication Protocol (PCEP) for fast protecting the ingress nodes of two types of paths or tunnels, which are Segment Routing (SR) paths and Bit Index Explicit Replication Tree/Traffic Engineering (BIER-TE) paths. The extensions comprise a foundation for protecting the ingress nodes of different types of paths. Based on this, the ingress protection of a new type of paths can be easily supported.

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 17 May 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
1.1. Terminologies	3
2. Path Ingress Protection Examples	4
2.1. SR Path Ingress Protection Example	4
2.2. BIER-TE Path Ingress Protection Example	5
3. Behavior around Ingress Failure	6
3.1. Source Detect	6
3.2. Backup Ingress Detect	6
3.3. Both Detect	7
4. Extensions to PCEP	7
4.1. Capabilities for Ingress Protection	7
4.1.1. Capability for Ingress Protection with Backup Ingress	7
4.1.2. Capability for Ingress Protection with Traffic Source	9
4.2. Extensions for Backup Ingress and Traffic Source	10
4.2.1. Extensions for Backup Ingress	10
4.2.2. Extensions for Traffic Source	16
5. Security Considerations	19
6. Acknowledgements	19
7. IANA Considerations	19
8. References	19
8.1. Normative References	19
8.2. Informative References	19
Authors' Addresses	20

1. Introduction

The fast protection of a transit node in each type of paths or tunnels have been proposed. For example, the fast protection of a transit node in a Segment Routing (SR) path or tunnel is described in [I-D.ietf-rtgwg-segment-routing-ti-lfa]. 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 an SR path/tunnel, a BIER-TE path/tunnel, or other type of paths/tunnels.

This document fills that void and specifies protocol extensions to Path Computation Element (PCE) communication Protocol (PCEP) [RFC5440] and [RFC9050] for fast protecting the ingress nodes of two types of paths: SR paths and BIER-TE paths. The extensions comprise a foundation for protecting the ingress nodes of different types of paths. Based on this, the ingress protection of a new type of paths can be easily supported.

Ingress node and ingress, fast protection and protection, path ingress protection and ingress protection, SR path and SR tunnel, as well as BIER-TE path and BIER-TE tunnel will be used exchangeably in the following sections.

1.1. Terminologies

The following terminologies are used in this document.

PCE: Path Computation Element or Path Computation Element server

PCEP: PCE communication Protocol

PCC: Path Computation Client

BIER: Bit Index Explicit Replication

BIFT: Bit Index Forwarding Table

CE: Customer Edge

PE: Provider Edge

TE: Traffic Engineering

SR: Segment Routing
 LFA: Loop-Free Alternate
 TI-LFA: Topology Independent LFA
 BFD: Bidirectional Forwarding Detection
 VPN: Virtual Private Network
 L3VPN: Layer 3 VPN
 FIB: Forwarding Information Base

2. Path Ingress Protection Examples

This section shows two examples of path ingress protection. One is SR path ingress protection, and the other is BIER-TE path ingress protection.

2.1. SR Path Ingress Protection Example

Figure 1 shows an example of protecting ingress PE1 of a SR path, which is from ingress PE1 to egress PE3 and represented by *** in the figure.

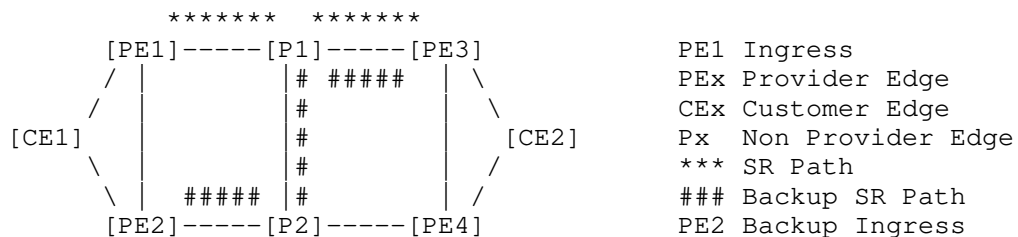


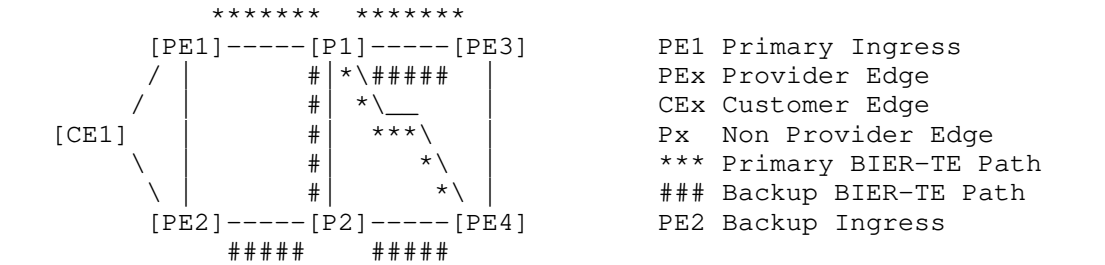
Figure 1: Protecting Ingress PE1 of SR Path

In normal operations, CE1 sends the traffic with destination PE3 to ingress PE1, which imports the traffic into the SR path.

When CE1 detects the failure of ingress PE1, it switches the traffic to backup ingress PE2, which imports the traffic from CE1 into a backup SR path. The backup path is from the backup ingress PE2 to the egress PE3 and represented by ### in the figure. When the traffic is imported into the backup path, it is sent to the egress PE3 along the path.

2.2. BIER-TE Path Ingress Protection Example

Figure 2 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.



When the PCC of the traffic source receives the information about the backup ingress, the primary ingress and the traffic, 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.

3. 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.

3.1. Source Detect

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

3.2. Backup Ingress Detect

The traffic source (e.g., CE1) always sends the traffic to both the ingress (e.g., PE1) of the primary path such as 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 path such as the backup BIER-TE path in normal operations. When it detects the failure of the ingress of the primary path, it imports the traffic from the source into the backup 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 path, it SHOULD consider this.

3.3. 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 path such as 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 path such as the backup BIER-TE path in normal operations. When it detects the failure of the ingress of the primary path, it imports the traffic from the source into the backup path.

4. Extensions to PCEP

A PCC runs on each of the edge nodes such as PEs of a network normally. A PCE runs on a server as a controller to communicate with PCCs. PCE and PCCs work together to support protection for the ingress of a path. The path is a SR path, a BIER-TE path, or a path of another type.

4.1. Capabilities for Ingress Protection

4.1.1. 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 each of different types of paths.

A new sub-TLV called INGRESS_PROTECTION_CAPABILITY is defined. It is included in the PATH_SETUP_TYPE_CAPABILITY TLV with PST = TBD1 (suggested value 2 for path ingress protection) 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.

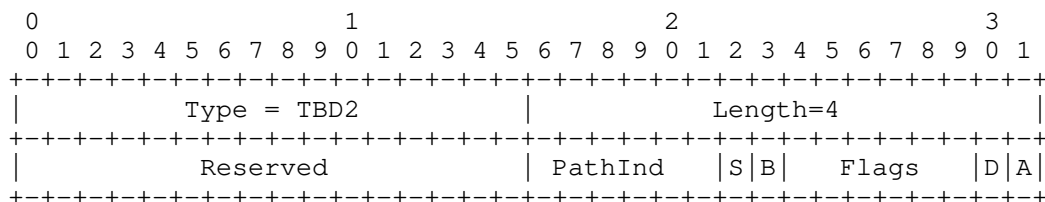


Figure 3: INGRESS_PROTECTION_CAPABILITY sub-TLV

Type: TBD2 is to be assigned by IANA.

Length: 4.

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

PathInd: 1 octet. Indicators for the types of paths whose ingress protections are supported. Two indicators are defined.

- o S : S = 1 indicating that the ingress protection of a SR path is supported.
- o B : B = 1 indicating that the ingress protection of a BIER-TE path is supported.

Flags: 1 octet. Two flags are defined.

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

A PCC, which supports ingress protection for different types of paths, sends a PCE an Open message containing INGRESS_PROTECTION_CAPABILITY sub-TLV. This sub-TLV indicates that the PCC is capable of supporting the ingress protection for the types of paths.

For example, if a PCC supports ingress protection for SR path and BIER-TE path, the PCC sends a PCE an Open message containing INGRESS_PROTECTION_CAPABILITY sub-TLV with S = 1 and B = 1.

A PCE, which supports ingress protection for different types of paths, sends a PCC an Open message containing INGRESS_PROTECTION_CAPABILITY sub-TLV. This sub-TLV indicates that the PCE is capable of supporting the ingress protection for the types of paths.

If both a PCC and a PCE support 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 an INGRESS_PROTECTION_CAPABILITY sub-TLV.

If a PCE receives an Open message from a PCC without a INGRESS_PROTECTION_CAPABILITY sub-TLV indicating PCC's support for the ingress protection of a type of paths, then the PCE MUST not send the PCC any request for ingress protection of the type of paths.

If a PCC receives an Open message from a PCE without a INGRESS_PROTECTION_CAPABILITY sub-TLV indicating PCE's support for the ingress protection of a type of paths, then the PCC MUST ignore any request for ingress protection of the type of paths 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 path, the PCC MUST let the forwarding entry for the backup path be Active.

4.1.2. 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 ingress protection.

The PCECC-CAPABILITY sub-TLV defined in [RFC9050] 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.

4.2. Extensions for Backup Ingress and Traffic Source

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 path, or

B2: import the traffic with possible service ID into the backup 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.

4.2.1. Extensions for Backup Ingress

For the packets from the traffic source, if the primary ingress (i.e., the ingress of the primary path) encapsulates the packets with a service ID or label into the path, the backup ingress MUST have this service ID or label and encapsulates the packets with the service ID or label into the backup 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 sub-TLV called INGRESS_PROTECTION is defined. When a PCE sends a PCC a PCInitiate message for initiating a backup path to protect the primary ingress node of a primary path, the message contains this TLV in the RP/SRP object. Its format is illustrated below.

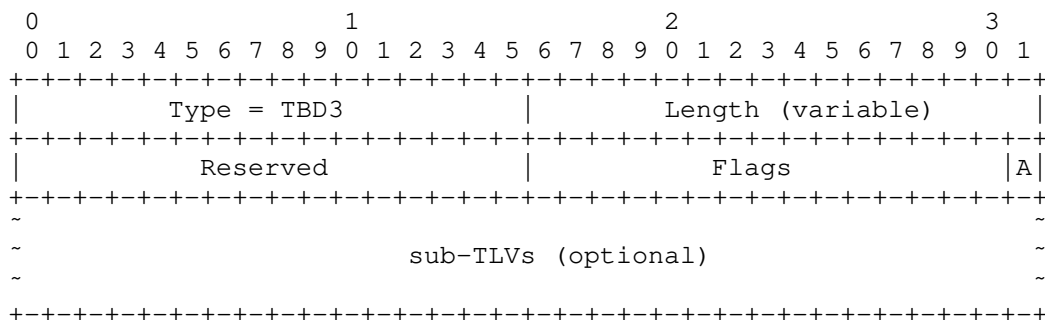


Figure 4: INGRESS_PROTECTION sub-TLV

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 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 path 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 path 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.

Three optional sub-TLVs are defined: Primary-Ingress sub-TLV, Service sub-TLV, and Traffic-Description sub-TLV. The Traffic-Description sub-TLV describes the traffic to be imported into the backup SR path. 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.

4.2.1.1. Primary-Ingress sub-TLV

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

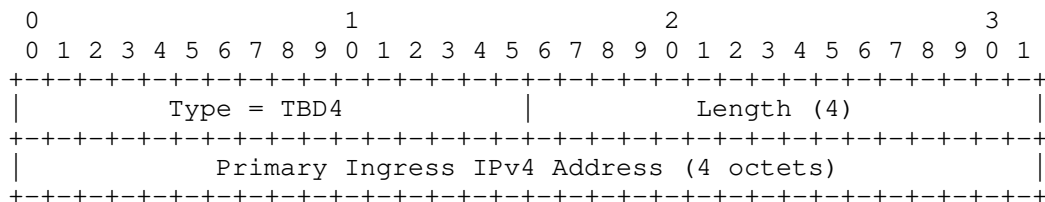


Figure 5: Primary Ingress IPv4 Address sub-TLV

Type: TBD4 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 path.

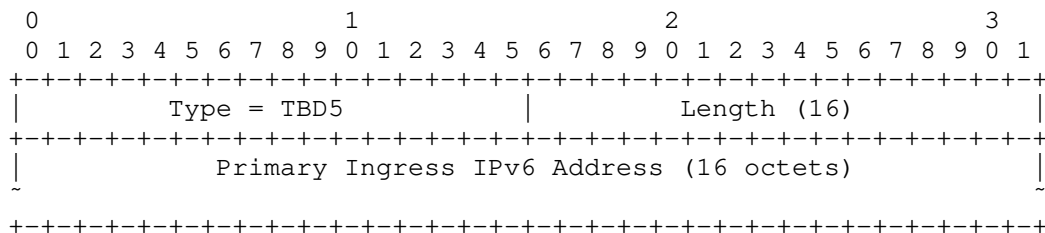


Figure 6: Primary Ingress IPv6 Address sub-TLV

Type: TBD5 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 path.

4.2.1.2. Service sub-TLV

A Service sub-TLV contains a service ID or label to be added into a packet to be carried by a path. 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.

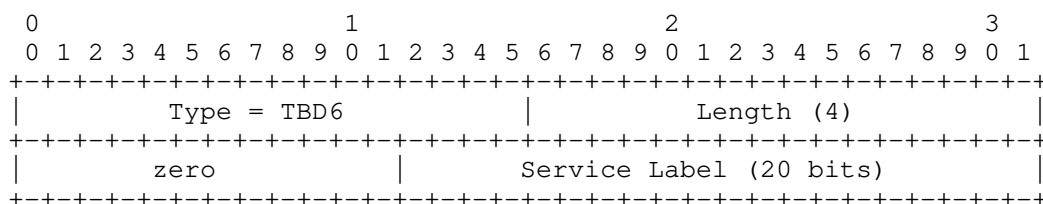


Figure 7: Service Label sub-TLV

Type: TBD6 is to be assigned by IANA.

Length: 4.

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

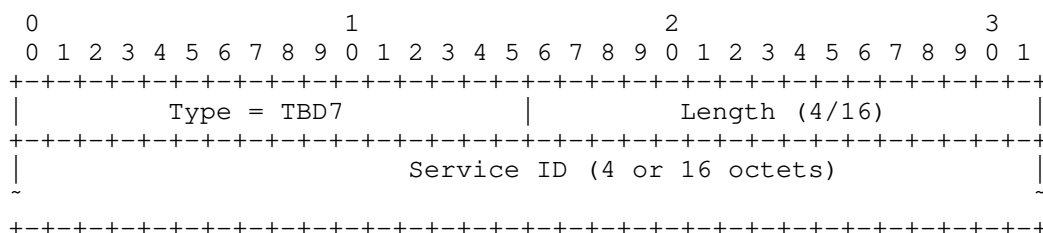


Figure 8: Service ID sub-TLV

Type: TBD7 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.

4.2.1.3. Traffic-Description sub-TLV

A Traffic-Description sub-TLV describes the traffic to be imported into a backup SR path. Its format is illustrated below.

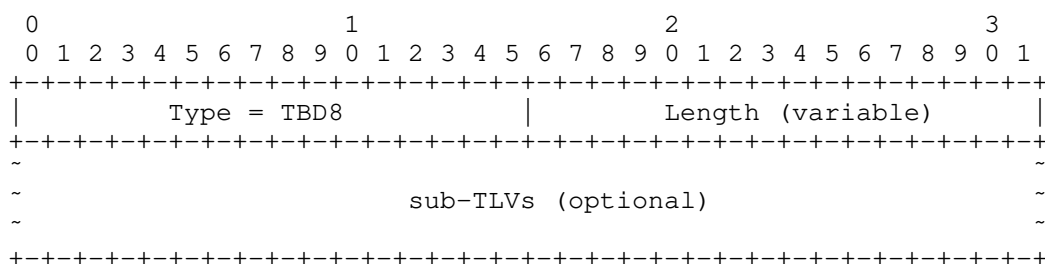


Figure 9: Traffic-Description sub-TLV

Type: TBD8 is to be assigned by IANA.

Length: Variable.

Two optional sub-TLVs are defined. One is FEC sub-TLV and the other interface sub-TLV.

A FEC sub-TLV describes the traffic to be imported into the backup path. It is an IP prefix with an optional virtual network ID. It has two formats: one for IPv4 and the other for IPv6, which are illustrated below.

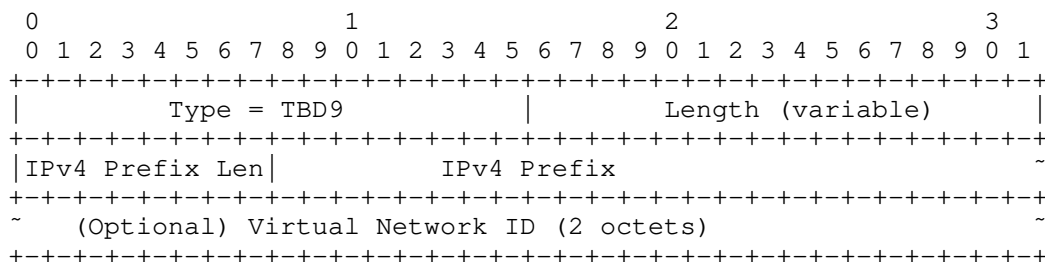


Figure 10: IPv4 FEC sub-TLV

Type: TBD9 is to be assigned by IANA.

Length: Variable.

IPv4 Prefix Len: Indicates the length of the IPv4 Prefix.

IPv4 Prefix: IPv4 Prefix rounded to octets.

Virtual Network ID: 2 octets. This is optional. It indicates the ID of a virtual network.

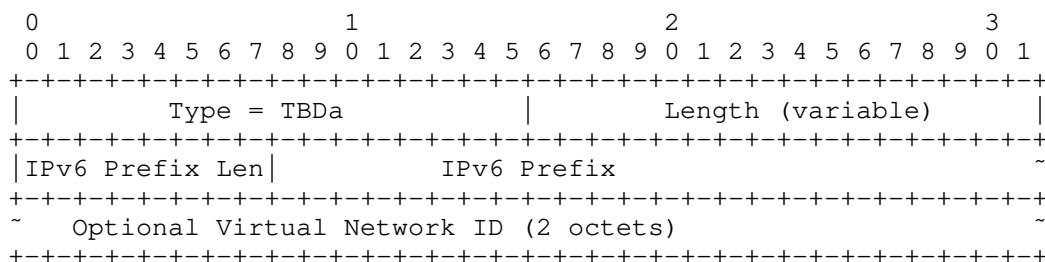


Figure 11: IPv6 FEC sub-TLV

Type: TBDA is to be assigned by IANA.

Length: Variable.

IPv6 Prefix Len: Indicates the length of the IPv6 Prefix.

IPv6 Prefix: IPv6 Prefix rounded to octets.

Virtual Network ID: 2 octets. This is optional. It indicates the ID of a virtual network.

An Interface sub-TLV indicates the interface from which the traffic is received and imported into the backup path. It has three formats: one for interface index, the other two for IPv4 and IPv6 address, which are illustrated below.

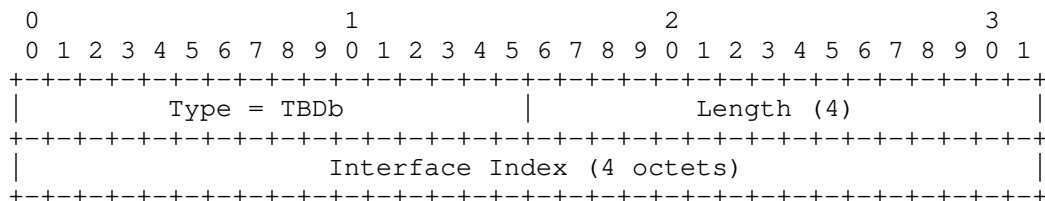


Figure 12: Interface Index sub-TLV

Type: TBDb is to be assigned by IANA.

Length: 4.

Interface Index: 4 octets. It indicates the index of an interface.

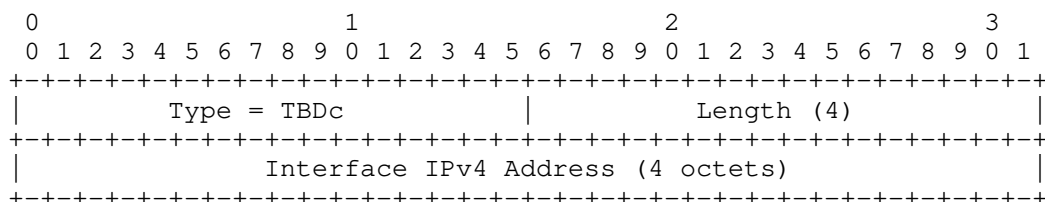


Figure 13: Interface IPv4 Address sub-TLV

Type: TBDc is to be assigned by IANA.

Length: 4.

Interface IPv4 Address: 4 octets. It represents the IPv4 address of an interface.

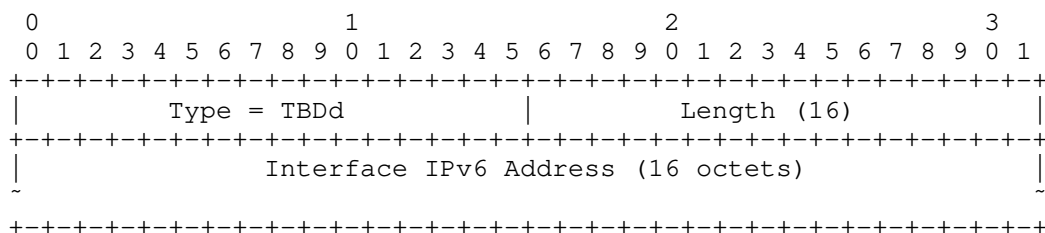


Figure 14: Interface IPv6 Address sub-TLV

Type: TBDd is to be assigned by IANA.

Length: 16.

Interface IPv6 Address: 16 octets. It represents the IPv6 address of an interface.

4.2.2. 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 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 [RFC9050] for a PCE as a controller to send instructions for LSPs to a PCC. This document defines a new object-type (TBDt) for 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.

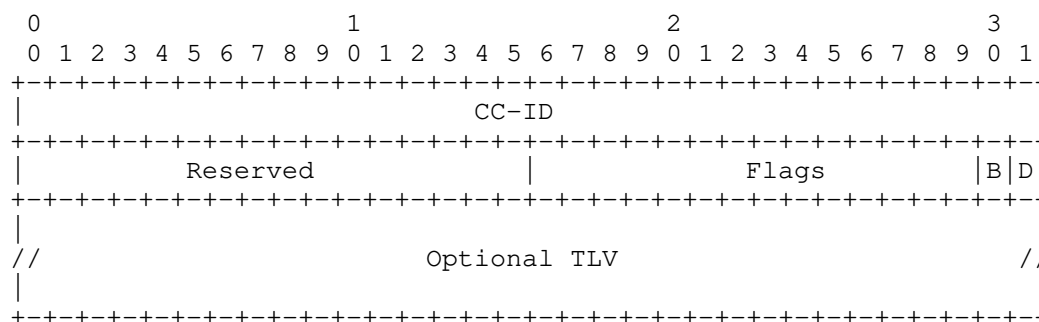


Figure 15: INGRESS-PROTECTION Object Body

CC-ID: It is the same as described in [RFC9050].

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, Traffic-Description TLV 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 Traffic-Description sub-TLV defined above is used as a TLV to contain the information about the traffic for a SR path 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 for a BIER-TE path in the object. A new TLV, called backup ingress TLV, is defined to contain the information about the backup ingress in the object.

4.2.2.1. Backup-Ingress TLV

A Backup-Ingress TLV indicates the IP address of the ingress node of a backup path. 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.

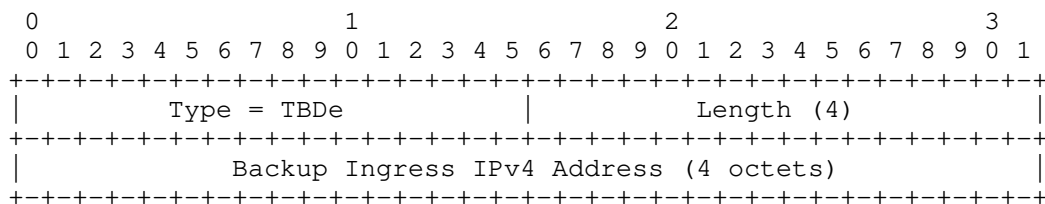


Figure 16: Backup Ingress IPv4 Address TLV

Type: TBDe is to be assigned by IANA.

Length: 4.

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

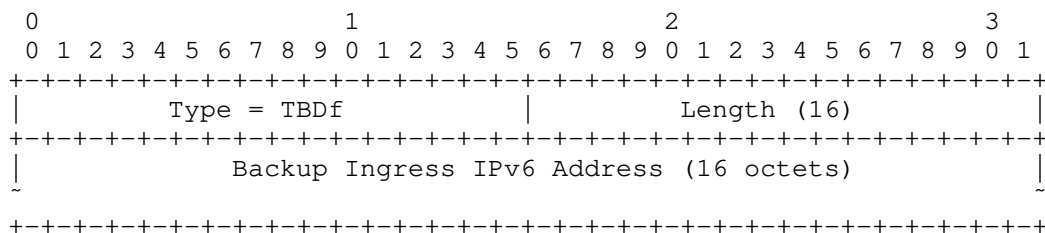


Figure 17: Backup Ingress IPv6 Address TLV

Type: TBDf 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.

5. Security Considerations

TBD

6. Acknowledgements

The authors of this document would like to thank Dhruv Dhody and Robin Li for their reviews and comments.

7. IANA Considerations

TBD

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>>.
- [RFC7356] Ginsberg, L., Previdi, S., and Y. Yang, "IS-IS Flooding Scope Link State PDUs (LSPs)", RFC 7356, DOI 10.17487/RFC7356, September 2014, <<https://www.rfc-editor.org/info/rfc7356>>.
- [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>>.
- [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>>.

8.2. 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-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>>.

[I-D.ietf-rtgwg-segment-routing-ti-lfa]

Litkowski, S., Bashandy, A., Filsfils, C., Francois, P., Decraene, B., and D. Voyer, "Topology Independent Fast Reroute using Segment Routing", Work in Progress, Internet-Draft, draft-ietf-rtgwg-segment-routing-ti-lfa-07, 29 June 2021, <<https://www.ietf.org/archive/id/draft-ietf-rtgwg-segment-routing-ti-lfa-07.txt>>.

[RFC5462] Andersson, L. and R. Asati, "Multiprotocol Label Switching (MPLS) Label Stack Entry: "EXP" Field Renamed to "Traffic Class" Field", RFC 5462, DOI 10.17487/RFC5462, February 2009, <<https://www.rfc-editor.org/info/rfc5462>>.

Authors' Addresses

Huaimo Chen
Futurewei
Boston, MA,
United States of America

Email: Huaimo.chen@futurewei.com

Mike McBride
Futurewei

Email: michael.mcbride@futurewei.com

Mehmet Toy
Verizon Inc.
United States of America

Email: mehmet.toy@verizon.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

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

Email: wangaj3@chinatelecom.cn

Zhenqiang Li
China Mobile
32 Xuanwumen West Ave, Xicheng District
Beijing
100053
China

Email: lizhengqiang@chinamobile.com

Yisong Liu
China Mobile

Email: liuyisong@chinamobile.com

Boris Khasanov
Yandex LLC
Moscow

Email: bhassanov@yahoo.com

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 Working Group
Internet-Draft
Updates: 5440 (if approved)
Intended status: Standards Track
Expires: 22 April 2022

D. Dhody
Huawei Technologies
19 October 2021

Updated Rules for PCEP Object Ordering
draft-dhody-pce-pcep-object-order-00

Abstract

The Path Computation Element Communication Protocol (PCEP) defines the mechanisms for the communication between a Path Computation Client (PCC) and a Path Computation Element (PCE), or among PCEs. Such interactions include include path computation requests and path computation replies defined in RFC 5440. As per RFC 5440, these message are required to follow strict object ordering.

This document updates RFC 5440 by relaxing the strict object ordering requirement in the PCEP messages.

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 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	2
2. Conventions	3
3. Motivation	3
4. Update to RFC 5440	3
5. Compatibility Considerations	3
6. Management Considerations	4
7. Other Efforts	4
8. Security Considerations	4
9. IANA Considerations	4
10. References	4
10.1. Normative References	4
10.2. Informative References	5
Appendix A. Acknowledgments	5
Appendix B. Examples	5
Author's Address	6

1. Introduction

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

[RFC5440] defines several PCEP messages. For each PCEP message type, rules are defined that specify the set of objects that the message can carry using [RFC5511]. Further, [RFC5440] states that the object ordering is mandatory. This causes confusion when multiple extensions add new objects in the PCEP messages and the respective order of these new objects is not specified (see [EID6627]).

This document updates [RFC5440] to relax the strict object ordering requirement.

2. Conventions

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. Motivation

The mandatory object ordering requirement in [RFC5440] is shown to result in exponential complexity in terms of what each new PCEP extension needs to cope with in terms of reconciling all previously-published RFCs, and all concurrently work in progress internet drafts. This requirement does not lend itself for extensibility of PCEP.

4. Update to RFC 5440

Section 6 of [RFC5440] states:

An implementation MUST form the PCEP messages using the object ordering specified in this document.

This text is updated to read as follows:

An implementation SHOULD form the PCEP messages using the object ordering specified in this and subsequent documents when an ordering can be unambiguously determined; an implementation MUST be prepared to receive a PCEP message with objects in any order.

This update does not aim to take away the object ordering completely. It is expected that the PCEP speaker will follow the object order as specified unless there are valid reasons to ignore. It is also expected that the receiver is able to unambiguously understand the object meaning irrespective of the order.

TODO: Scan all PCEP extensions to see if any other text needs to be updated related to object ordering.

5. Compatibility Considerations

While one of the main objectives of the changes made by this document is to enable backward compatibility between PCEP extensions, there remains an issue of compatibility between existing implementations of [RFC5440] and implementations that are consistent with this document.

It should be noted that common behavior for checking object ordering in received PCEP messages is as described by the updated text presented in Section 4. Thus, many implementations, will still have implemented a consistent and future-proof approach. However, for completeness, it is worth noting how behaviors might interact between implementations.

The messages generated by an implementation of this document when received by a legacy implementation with a strict interpretation of object ordering MAY lead to error handling. It is interesting to note that the [RFC5440] does not define an Error-Type and Error-value corresponding to this error condition.

6. Management Considerations

Implementations receiving set objects that they consider out of order MAY log this. That could be helpful for diagnosing backward compatibility issues.

7. Other Efforts

In the past there have been effort to consolidate and update the RBNF such as in [I-D.cmfg-pce-pcep-grammar]. This document document relaxes the object ordering only, it does not take on the various other issues or the need to consolidate the RBNF for all PCEP extensions. They might be taken up in parallel.

8. Security Considerations

This document does not raise any security issues.

9. IANA Considerations

This document does not require any IANA actions.

10. References

10.1. Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <<https://www.rfc-editor.org/info/rfc2119>>.
- [RFC5440] Vasseur, JP., Ed. and JL. Le Roux, Ed., "Path Computation Element (PCE) Communication Protocol (PCEP)", RFC 5440, DOI 10.17487/RFC5440, March 2009, <<https://www.rfc-editor.org/info/rfc5440>>.

- [RFC5511] Farrel, A., "Routing Backus-Naur Form (RBNF): A Syntax Used to Form Encoding Rules in Various Routing Protocol Specifications", RFC 5511, DOI 10.17487/RFC5511, April 2009, <<https://www.rfc-editor.org/info/rfc5511>>.
- [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>>.

10.2. Informative References

- [EID6627] "Errata ID: 6627", n.d., <<https://www.rfc-editor.org/errata/eid6627>>.
- [I-D.cmfg-pce-pcep-grammar]
Casellas, R., Margaria, C., Farrel, A., Dios, O. G. D., Dhody, D., and X. Zhang, "Current issues with existing RBNF notation for PCEP messages and extensions", Work in Progress, Internet-Draft, draft-cmfg-pce-pcep-grammar-02, 10 January 2014, <<https://www.ietf.org/archive/id/draft-cmfg-pce-pcep-grammar-02.txt>>.
- [RFC5455] Sivabalan, S., Ed., Parker, J., Boutros, S., and K. Kumaki, "Diffserv-Aware Class-Type Object for the Path Computation Element Communication Protocol", RFC 5455, DOI 10.17487/RFC5455, March 2009, <<https://www.rfc-editor.org/info/rfc5455>>.
- [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>>.

Appendix A. Acknowledgments

Thanks to John Scudder for the motivation behind this document.
Thanks to Oscar Gonzalez de Dios and Cyril Margaria for raising errata on this topic. Thanks to the author of [I-D.cmfg-pce-pcep-grammar] for highlighting the issue.

Appendix B. Examples

As described in [EID6627], for the PCReq message, the CLASSTYPE object is encoded after the END-POINTS object in [RFC5455]. Where as in [RFC8231], the LSP object is encoded just after the END-POINTS object. So it is not known which of the below order is expected.

...<END-POINTS>[<LSP>][<CLASSTYPE>]...

or

...<END-POINTS>[<CLASSTYPE>][<LSP>]...

This update require the receiver to be able to except both of these.

Author's Address

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

Email: dhruv.ietf@gmail.com

PCE Working Group
Internet-Draft
Updates: 5440 (if approved)
Intended status: Standards Track
Expires: 6 September 2022

D. Dhody
Huawei Technologies
5 March 2022

Updated Rules for PCEP Object Ordering
draft-dhody-pce-pcep-object-order-01

Abstract

The Path Computation Element Communication Protocol (PCEP) defines the mechanisms for the communication between a Path Computation Client (PCC) and a Path Computation Element (PCE), or among PCEs. Such interactions include include path computation requests and path computation replies defined in RFC 5440. As per RFC 5440, these message are required to follow strict object ordering.

This document updates RFC 5440 by relaxing the strict object ordering requirement in the PCEP messages.

Status of This Memo

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

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

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

This Internet-Draft will expire on 6 September 2022.

Copyright Notice

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

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

and restrictions with respect to this document. Code Components extracted from this document must include Revised BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Revised BSD License.

Table of Contents

1. Introduction	2
2. Conventions	3
3. Motivation	3
4. Update to RFC 5440	3
5. Compatibility Considerations	3
6. Open Questions	4
7. Management Considerations	4
8. Other Efforts	4
9. Security Considerations	4
10. IANA Considerations	4
11. References	4
11.1. Normative References	4
11.2. Informative References	5
Appendix A. Acknowledgments	5
Appendix B. Examples	6
Appendix C. When Order Matters	6
Author's Address	6

1. Introduction

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

[RFC5440] defines several PCEP messages. For each PCEP message type, rules are defined that specify the set of objects that the message can carry using [RFC5511]. Further, [RFC5440] states that the object ordering is mandatory. This causes confusion when multiple extensions add new objects in the PCEP messages and the respective order of these new objects is not specified (see [EID6627]).

This document updates [RFC5440] to relax the strict object ordering requirement.

2. Conventions

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. Motivation

The mandatory object ordering requirement in [RFC5440] is shown to result in exponential complexity in terms of what each new PCEP extension needs to cope with in terms of reconciling all previously-published RFCs, and all concurrently work in progress internet drafts. This requirement does not lend itself for extensibility of PCEP.

4. Update to RFC 5440

Section 6 of [RFC5440] states:

An implementation **MUST** form the PCEP messages using the object ordering specified in this document.

This text is updated to read as follows:

An implementation **SHOULD** form the PCEP messages using the object ordering specified in this and subsequent documents when an ordering can be unambiguously determined; an implementation **MUST** be prepared to receive a PCEP message with objects in any order.

This update does not aim to take away the object ordering completely. It is expected that the PCEP speaker will follow the object order as specified unless there are valid reasons to ignore. It is also expected that the receiver is able to unambiguously understand the object meaning irrespective of the order.

5. Compatibility Considerations

While one of the main objectives of the changes made by this document is to enable backward compatibility between PCEP extensions, there remains an issue of compatibility between existing implementations of [RFC5440] and implementations that are consistent with this document.

It should be noted that common behavior for checking object ordering in received PCEP messages is as described by the updated text presented in Section 4. Thus, many implementations, will still have

implemented a consistent and future-proof approach. However, for completeness, it is worth noting how behaviors might interact between implementations.

The messages generated by an implementation of this document when received by a legacy implementation with a strict interpretation of object ordering MAY lead to error handling. It is interesting to note that the [RFC5440] does not define an Error-Type and Error-value corresponding to this error condition.

6. Open Questions

- * Should a new flag or a TLV in Open Message be added to exchange this capability? Not sure if this is strictly needed.

7. Management Considerations

Implementations receiving set objects that they consider out of order MAY log this. That could be helpful for diagnosing backward compatibility issues.

8. Other Efforts

In the past there have been effort to consolidate and update the RBNF such as in [I-D.cmfg-pce-pcep-grammar]. This document document relaxes the object ordering only, it does not take on the various other issues or the need to consolidate the RBNF for all PCEP extensions. There have been proposal to consolidate the RBNF for the PCEP message in a single place in GitHub and use modern data modeling tools to represent PCEP extensions. They might be taken up in parallel.

9. Security Considerations

This document does not raise any security issues.

10. IANA Considerations

This document does not require any IANA actions.

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>>.

- [RFC5440] Vasseur, JP., Ed. and JL. Le Roux, Ed., "Path Computation Element (PCE) Communication Protocol (PCEP)", RFC 5440, DOI 10.17487/RFC5440, March 2009, <<https://www.rfc-editor.org/info/rfc5440>>.
- [RFC5511] Farrel, A., "Routing Backus-Naur Form (RBNF): A Syntax Used to Form Encoding Rules in Various Routing Protocol Specifications", RFC 5511, DOI 10.17487/RFC5511, April 2009, <<https://www.rfc-editor.org/info/rfc5511>>.
- [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>>.

11.2. Informative References

- [EID6627] "Errata ID: 6627", n.d., <<https://www.rfc-editor.org/errata/eid6627>>.
- [I-D.cmfg-pce-pcep-grammar]
Casellas, R., Margaria, C., Farrel, A., Dios, O. G. D., Dhody, D., and X. Zhang, "Current issues with existing RBNF notation for PCEP messages and extensions", Work in Progress, Internet-Draft, draft-cmfg-pce-pcep-grammar-02, 10 January 2014, <<https://www.ietf.org/archive/id/draft-cmfg-pce-pcep-grammar-02.txt>>.
- [RFC5455] Sivabalan, S., Ed., Parker, J., Boutros, S., and K. Kumaki, "Diffserv-Aware Class-Type Object for the Path Computation Element Communication Protocol", RFC 5455, DOI 10.17487/RFC5455, March 2009, <<https://www.rfc-editor.org/info/rfc5455>>.
- [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>>.

Appendix A. Acknowledgments

Thanks to John Scudder for the motivation behind this document.
Thanks to Oscar Gonzalez de Dios and Cyril Margaria for raising errata on this topic. Thanks to the author of [I-D.cmfg-pce-pcep-grammar] for highlighting the issue.

Appendix B. Examples

As described in [EID6627], for the PCReq message, the CLASSTYPE object is encoded after the END-POINTS object in [RFC5455]. Where as in [RFC8231], the LSP object is encoded just after the END-POINTS object. So it is not known which of the below order is expected.

...<END-POINTS>[<LSP>][<CLASSTYPE>]...

or

...<END-POINTS>[<CLASSTYPE>][<LSP>]...

This update require the receiver to be able to except both of these.

Appendix C. When Order Matters

There are cases where the ordering between objects is important. For instance PCRpt message [RFC8231] includes <path> with some attributes say BANDWIDTH can be part of both <actual-attribute-list> and <intended-attribute-list>.

Where:

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

An important factor to distinguish between the actual and intended attribute list is the presence of RRO (i.e. <actual-path>) and the order of objects in the PCRpt message.

If the RRO is present, any attributes encoded before it, are to be considered as part of <actual-attribute-list> and those after it, as part of <intended-attribute-list>.

If the RRO is absent, all attributes are part of <intended-attribute-list>.

Thus the approach taken by this document is to say that ordering is relaxed in cases where there is no ambiguity.

Author's Address

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

Email: dhruv.ietf@gmail.com

PCE Working Group
Internet-Draft
Intended status: Experimental
Expires: 26 February 2022

D. Dhody
S. Peng
Huawei Technologies
Y. Lee
Samsung Electronics
D. Ceccarelli
Ericsson
A. Wang
China Telecom
G. Mishra
Verizon Inc.
S. Sivabalan
Ciena Corporation
25 August 2021

PCEP extensions for Distribution of Link-State and TE Information
draft-dhodylee-pce-pcep-ls-22

Abstract

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

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

Requirements Language

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

Status of This Memo

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

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

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

This Internet-Draft will expire on 26 February 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	4
1.1. Scope	5
2. Terminology	6
3. Applicability	6
4. Requirements for PCEP extensions	7
5. New Functions to distribute link-state (and TE) via PCEP . .	8
6. Overview of Extensions to PCEP	9
6.1. New Messages	9
6.2. Capability Advertisement	9
6.3. Initial Link-State (and TE) Synchronization	10
6.3.1. Optimizations for LS Synchronization	12
6.4. LS Report	12
7. Transport	12
8. PCEP Messages	13
8.1. LS Report Message	13
8.2. The PCErr Message	13
9. Objects and TLV	14
9.1. TLV Format	14
9.2. Open Object	14
9.2.1. LS Capability TLV	14
9.3. LS Object	15
9.3.1. Routing Universe TLV	17
9.3.2. Route Distinguisher TLV	18
9.3.3. Virtual Network TLV	18
9.3.4. Local Node Descriptors TLV	18

9.3.5. Remote Node Descriptors TLV	19
9.3.6. Node Descriptors Sub-TLVs	20
9.3.7. Link Descriptors TLV	21
9.3.8. Prefix Descriptors TLV	21
9.3.9. PCEP-LS Attributes	22
9.3.9.1. Node Attributes TLV	22
9.3.9.2. Link Attributes TLV	22
9.3.9.3. Prefix Attributes TLV	23
9.3.10. Removal of an Attribute	23
10. Other Considerations	24
10.1. Inter-AS Links	24
11. Security Considerations	24
12. Manageability Considerations	24
12.1. Control of Function and Policy	24
12.2. Information and Data Models	25
12.3. Liveness Detection and Monitoring	25
12.4. Verify Correct Operations	25
12.5. Requirements On Other Protocols	26
12.6. Impact On Network Operations	26
13. IANA Considerations	26
13.1. PCEP Messages	26
13.2. PCEP Objects	26
13.3. LS Object	26
13.4. PCEP-Error Object	27
13.5. PCEP TLV Type Indicators	28
13.6. PCEP-LS Sub-TLV Type Indicators	28
14. TLV Code Points Summary	29
15. Implementation Status	30
15.1. Hierarchical Transport PCE controllers	30
15.2. ONOS-based Controller (MDSC and PNC)	31
16. Acknowledgments	31
17. References	31
17.1. Normative References	31
17.2. Informative References	32
Appendix A. Examples	35
A.1. All Nodes	36
A.2. Designated Node	37
A.3. Between PCEs	37
Appendix B. Contributor Addresses	38
Authors' Addresses	39

1. Introduction

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

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

Section 4.3 of [RFC4655] describes the potential load of the TED on a network node and proposes an architecture where the TED is maintained by the PCE rather than the network nodes. However, it does not describe how a PCE would obtain the information needed to populate its TED. PCE may construct its TED by participating in the IGP ([RFC3630] and [RFC5305] for MPLS-TE; [RFC4203] and [RFC5307] for GMPLS). An alternative mechanism is offered by BGP-LS [I-D.ietf-idr-rfc7752bis] .

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

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

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

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

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

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

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

1.1. Scope

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

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

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

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

2. Terminology

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

3. Applicability

The mechanism specified in this draft is applicable to deployments:

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

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

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

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

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

4. Requirements for PCEP extensions

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

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

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

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

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

6. Overview of Extensions to PCEP

6.1. New Messages

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

6.2. Capability Advertisement

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

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

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

6.3. Initial Link-State (and TE) Synchronization

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

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

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

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

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

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

The successful LS Synchronization sequence is shown in Figure 1.

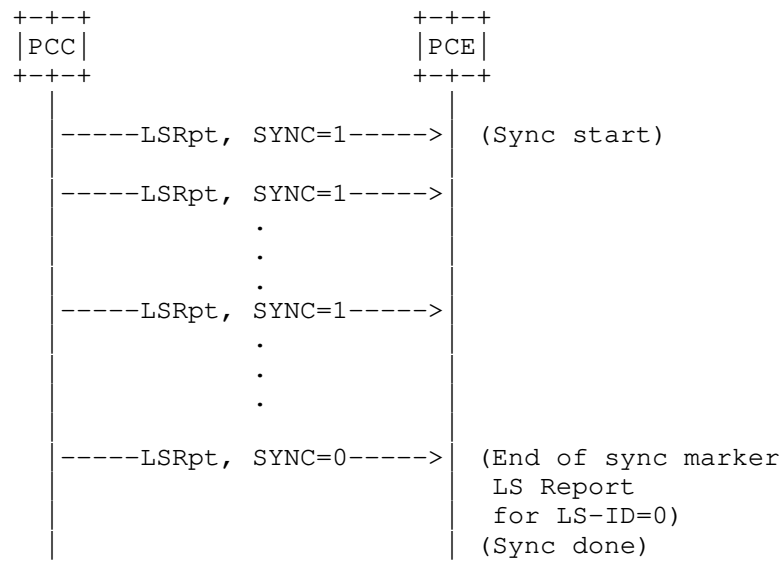


Figure 1: Successful LS synchronization

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

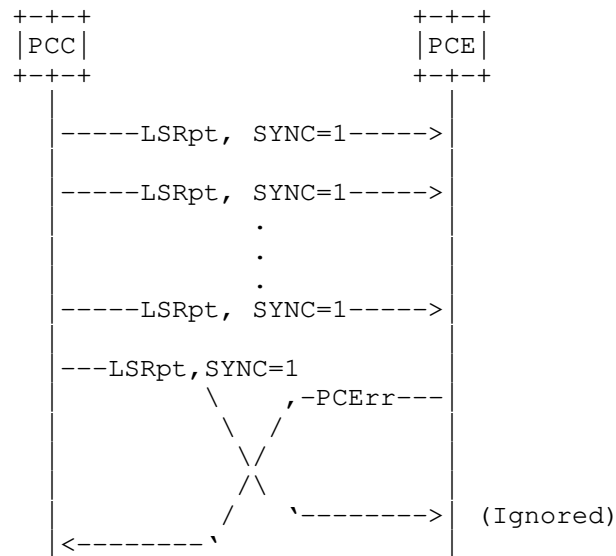


Figure 2: Failed LS synchronization (PCE failure)

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

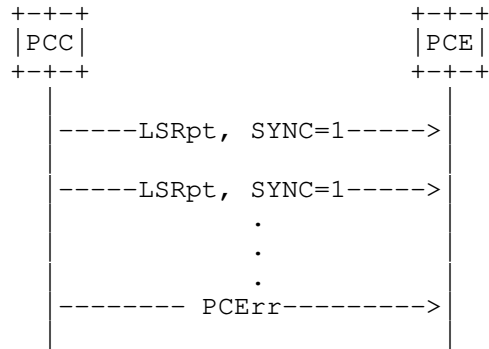


Figure 3: Failed LS synchronization (PCC failure)

6.3.1. Optimizations for LS Synchronization

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

6.4. LS Report

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

More details about the LSRpt message are in Section 8.1.

7. Transport

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

8. PCEP Messages

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

8.1. LS Report Message

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

The format of the LSRpt message is as follows:

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

Where:

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

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

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

8.2. The PCErr Message

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

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

```

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

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

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

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

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

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

```

9. Objects and TLV

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

9.1. TLV Format

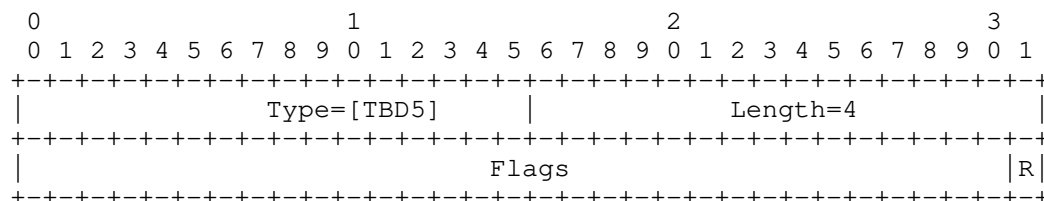
The TLV and the sub-TLV format (and padding) in this document, is as per section 7.1 of [RFC5440].

9.2. Open Object

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

9.2.1. LS Capability TLV

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



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

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

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

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

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

9.3. LS Object

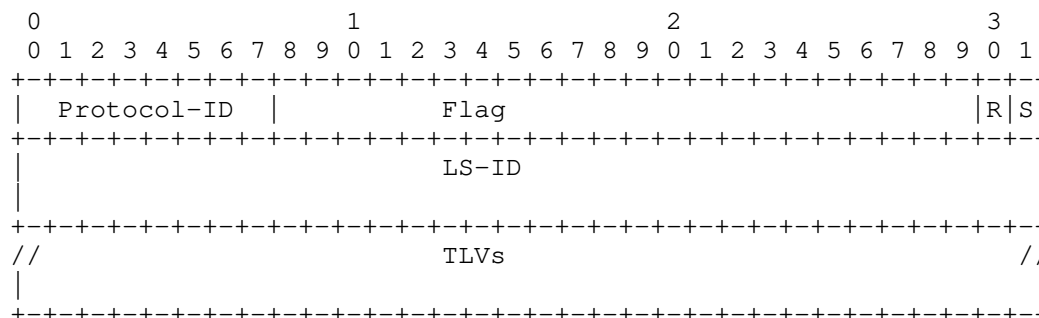
The LS (link-state) object MUST be carried within LSRpt messages and MAY be carried within PCErr messages. The LS object contains a set of fields used to specify the target node or link. It also contains a flag indicating to a PCE that the LS synchronization is in progress. The TLVs used with the LS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

LS Object-Class is TBD6.

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

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

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



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

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

Flags (24-bit):

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

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

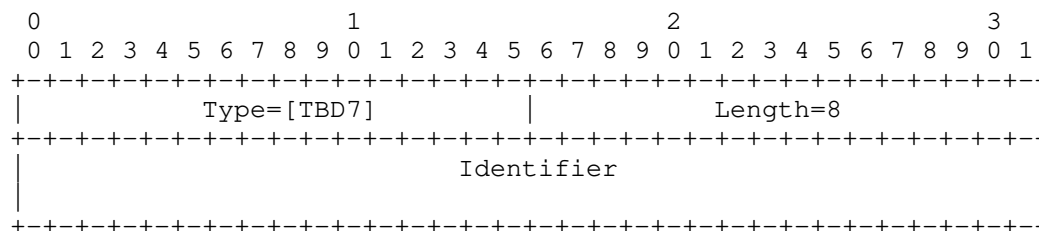
Unassigned bits are considered reserved. They MUST be set to 0 on transmission and MUST be ignored on receipt.

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

9.3.1. Routing Universe TLV

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

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



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

Identifier	Routing Universe
0	Default Layer 3 Routing topology

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

9.3.2. Route Distinguisher TLV

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

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

9.3.3. Virtual Network TLV

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

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

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

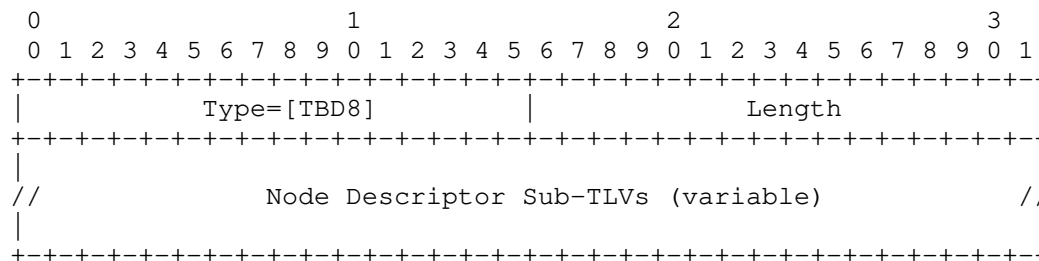
9.3.4. Local Node Descriptors TLV

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

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

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

This TLV is applicable for all LS Object-Type.

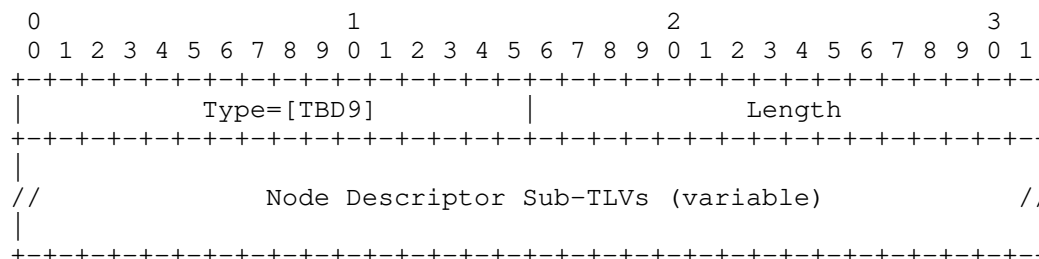


The value contains one or more Node Descriptor Sub-TLVs defined in Section 9.3.6.

9.3.5. Remote Node Descriptors TLV

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

This TLV is applicable for LS Link Object-Type.



9.3.6. Node Descriptors Sub-TLVs

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

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

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

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

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

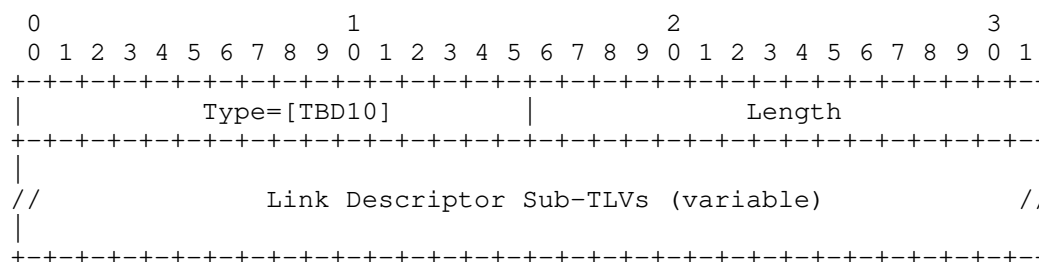
A new sub-TLV type (1) is allocated for SPEAKER-ENTITY-ID sub-TLV. The length and value fields are as per [RFC8232].

9.3.7. Link Descriptors TLV

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

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

This TLV is applicable for LS Link Object-Type.



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

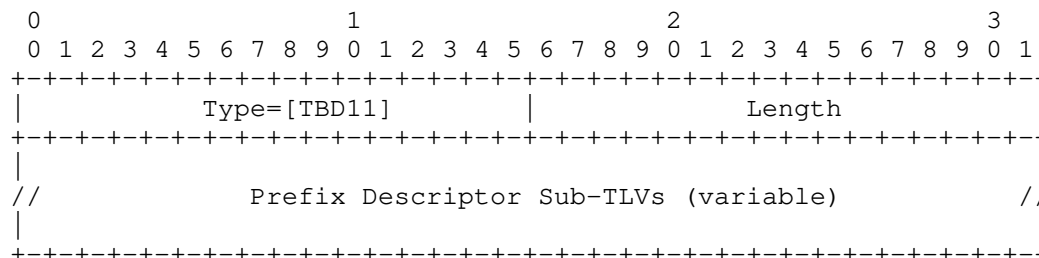
The format and semantics of the 'value' fields in most 'Link Descriptor' sub-TLVs correspond to the format and semantics of value fields in IS-IS Extended IS Reachability sub-TLVs, defined in [RFC5305], [RFC5307] and [RFC6119]. Although the encodings for 'Link Descriptor' TLVs were originally defined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

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

9.3.8. Prefix Descriptors TLV

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

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

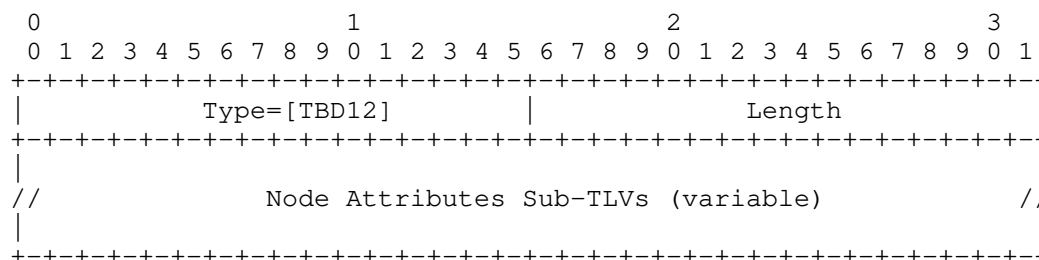


All Prefix Descriptors TLVs defined for BGP-LS can then be used with PCEP-LS as well. No new PCEP sub-TLVs for Prefix Descriptor are defined in this document.

9.3.9. PCEP-LS Attributes

9.3.9.1. Node Attributes TLV

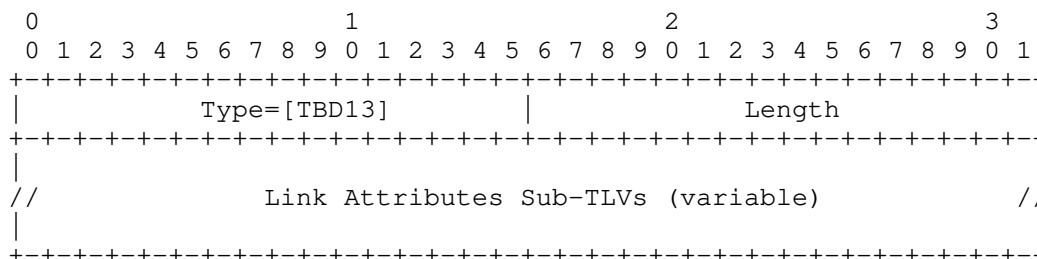
This is an optional attribute that is used to carry node attributes. This TLV is applicable for LS Node Object-Type.



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

9.3.9.2. Link Attributes TLV

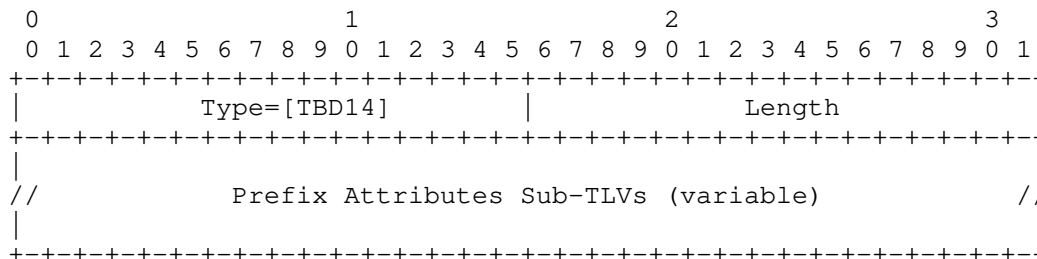
This TLV is applicable for LS Link Object-Type. The format and semantics of the 'value' fields in some 'Link Attribute' sub-TLVs correspond to the format and semantics of the 'value' fields in IS-IS Extended IS Reachability sub-TLVs, defined in [RFC5305], [RFC5307] and [I-D.ietf-idr-rfc7752bis]. Although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



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

9.3.9.3. Prefix Attributes TLV

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



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

9.3.10. Removal of an Attribute

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

10. Other Considerations

10.1. Inter-AS Links

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

11. Security Considerations

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

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

12. Manageability Considerations

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

12.1. Control of Function and Policy

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

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

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

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

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

12.2. Information and Data Models

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

An implementation SHOULD also provide the statistics:

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

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

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

12.3. Liveness Detection and Monitoring

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

12.4. Verify Correct Operations

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

12.5. Requirements On Other Protocols

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

12.6. Impact On Network Operations

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

13. IANA Considerations

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

13.1. PCEP Messages

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

Value	Meaning	Reference
TBD3	LSRpt	[This I-D]

13.2. PCEP Objects

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

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

13.3. LS Object

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

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

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

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

The following values are defined in this document:

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

13.4. PCEP-Error Object

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

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

13.5. PCEP TLV Type Indicators

This document defines the following new PCEP TLVs.

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

13.6. PCEP-LS Sub-TLV Type Indicators

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

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

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

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

Value	Meaning
1	SPEAKER-ENTITY-ID

14. TLV Code Points Summary

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

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

* this TLV is defined in a different PCEP document

Figure 4: TLV Table

15. Implementation Status

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

15.1. Hierarchical Transport PCE controllers

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

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

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

15.2. ONOS-based Controller (MDSC and PNC)

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

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

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

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

16. Acknowledgments

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

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

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

17. References

17.1. Normative References

[I-D.ietf-idr-rfc7752bis]
Talaulikar, K., "Distribution of Link-State and Traffic Engineering Information Using BGP", Work in Progress, Internet-Draft, draft-ietf-idr-rfc7752bis-08, 26 July 2021, <<https://datatracker.ietf.org/doc/html/draft-ietf-idr-rfc7752bis-08>>.

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <<https://www.rfc-editor.org/info/rfc2119>>.
- [RFC5305] Li, T. and H. Smit, "IS-IS Extensions for Traffic Engineering", RFC 5305, DOI 10.17487/RFC5305, October 2008, <<https://www.rfc-editor.org/info/rfc5305>>.
- [RFC5307] Kompella, K., Ed. and Y. Rekhter, Ed., "IS-IS Extensions in Support of Generalized Multi-Protocol Label Switching (GMPLS)", RFC 5307, DOI 10.17487/RFC5307, October 2008, <<https://www.rfc-editor.org/info/rfc5307>>.
- [RFC5440] Vasseur, JP., Ed. and JL. Le Roux, Ed., "Path Computation Element (PCE) Communication Protocol (PCEP)", RFC 5440, DOI 10.17487/RFC5440, March 2009, <<https://www.rfc-editor.org/info/rfc5440>>.
- [RFC6119] Harrison, J., Berger, J., and M. Bartlett, "IPv6 Traffic Engineering in IS-IS", RFC 6119, DOI 10.17487/RFC6119, February 2011, <<https://www.rfc-editor.org/info/rfc6119>>.
- [RFC8174] Leiba, B., "Ambiguity of Uppercase vs Lowercase in RFC 2119 Key Words", BCP 14, RFC 8174, DOI 10.17487/RFC8174, May 2017, <<https://www.rfc-editor.org/info/rfc8174>>.
- [RFC8232] Crabbe, E., Minei, I., Medved, J., Varga, R., Zhang, X., and D. Dhody, "Optimizations of Label Switched Path State Synchronization Procedures for a Stateful PCE", RFC 8232, DOI 10.17487/RFC8232, September 2017, <<https://www.rfc-editor.org/info/rfc8232>>.

17.2. Informative References

- [I-D.ietf-pce-pcep-flowspec] Dhody, D., Farrel, A., and Z. Li, "PCEP Extension for Flow Specification", Work in Progress, Internet-Draft, draft-ietf-pce-pcep-flowspec-12, 30 October 2020, <<https://datatracker.ietf.org/doc/html/draft-ietf-pce-pcep-flowspec-12>>.

- [I-D.ietf-pce-pcep-yang]
Dhody, D., Hardwick, J., Beeram, V. P., and J. Tantsura,
"A YANG Data Model for Path Computation Element
Communications Protocol (PCEP)", Work in Progress,
Internet-Draft, draft-ietf-pce-pcep-yang-16, 22 February
2021, <<https://datatracker.ietf.org/doc/html/draft-ietf-pce-pcep-yang-16>>.
- [I-D.ietf-pce-vn-association]
Lee, Y., Zheng, H., and D. Ceccarelli, "Path Computation
Element communication Protocol (PCEP) extensions for
Establishing Relationships between sets of LSPs and
Virtual Networks", Work in Progress, Internet-Draft,
draft-ietf-pce-vn-association-04, 16 April 2021,
<<https://datatracker.ietf.org/doc/html/draft-ietf-pce-vn-association-04>>.
- [I-D.ietf-teas-actn-requirements]
Lee, Y., Ceccarelli, D., Miyasaka, T., Shin, J. Y., and K.
Lee, "Requirements for Abstraction and Control of TE
Networks", Work in Progress, Internet-Draft, draft-ietf-
teas-actn-requirements-09, 2 March 2018,
<<https://datatracker.ietf.org/doc/html/draft-ietf-teas-actn-requirements-09>>.
- [I-D.kondreddy-pce-pcep-ls-sync-optimizations]
Kondreddy, V. R. and M. S. Negi, "Optimizations of PCEP
Link-State(LS) Synchronization Procedures", Work in
Progress, Internet-Draft, draft-kondreddy-pce-pcep-ls-
sync-optimizations-00, 9 October 2015,
<<https://datatracker.ietf.org/doc/html/draft-kondreddy-pce-pcep-ls-sync-optimizations-00>>.
- [RFC3630] Katz, D., Kompella, K., and D. Yeung, "Traffic Engineering
(TE) Extensions to OSPF Version 2", RFC 3630,
DOI 10.17487/RFC3630, September 2003,
<<https://www.rfc-editor.org/info/rfc3630>>.
- [RFC4203] Kompella, K., Ed. and Y. Rekhter, Ed., "OSPF Extensions in
Support of Generalized Multi-Protocol Label Switching
(GMPLS)", RFC 4203, DOI 10.17487/RFC4203, October 2005,
<<https://www.rfc-editor.org/info/rfc4203>>.
- [RFC4364] Rosen, E. and Y. Rekhter, "BGP/MPLS IP Virtual Private
Networks (VPNs)", RFC 4364, DOI 10.17487/RFC4364, February
2006, <<https://www.rfc-editor.org/info/rfc4364>>.

- [RFC4655] Farrel, A., Vasseur, J.-P., and J. Ash, "A Path Computation Element (PCE)-Based Architecture", RFC 4655, DOI 10.17487/RFC4655, August 2006, <<https://www.rfc-editor.org/info/rfc4655>>.
- [RFC5316] Chen, M., Zhang, R., and X. Duan, "ISIS Extensions in Support of Inter-Autonomous System (AS) MPLS and GMPLS Traffic Engineering", RFC 5316, DOI 10.17487/RFC5316, December 2008, <<https://www.rfc-editor.org/info/rfc5316>>.
- [RFC5392] Chen, M., Zhang, R., and X. Duan, "OSPF Extensions in Support of Inter-Autonomous System (AS) MPLS and GMPLS Traffic Engineering", RFC 5392, DOI 10.17487/RFC5392, January 2009, <<https://www.rfc-editor.org/info/rfc5392>>.
- [RFC5925] Touch, J., Mankin, A., and R. Bonica, "The TCP Authentication Option", RFC 5925, DOI 10.17487/RFC5925, June 2010, <<https://www.rfc-editor.org/info/rfc5925>>.
- [RFC6549] Lindem, A., Roy, A., and S. Mirtorabi, "OSPFv2 Multi-Instance Extensions", RFC 6549, DOI 10.17487/RFC6549, March 2012, <<https://www.rfc-editor.org/info/rfc6549>>.
- [RFC6805] King, D., Ed. and A. Farrel, Ed., "The Application of the Path Computation Element Architecture to the Determination of a Sequence of Domains in MPLS and GMPLS", RFC 6805, DOI 10.17487/RFC6805, November 2012, <<https://www.rfc-editor.org/info/rfc6805>>.
- [RFC6952] Jethanandani, M., Patel, K., and L. Zheng, "Analysis of BGP, LDP, PCEP, and MSDP Issues According to the Keying and Authentication for Routing Protocols (KARP) Design Guide", RFC 6952, DOI 10.17487/RFC6952, May 2013, <<https://www.rfc-editor.org/info/rfc6952>>.
- [RFC7525] Sheffer, Y., Holz, R., and P. Saint-Andre, "Recommendations for Secure Use of Transport Layer Security (TLS) and Datagram Transport Layer Security (DTLS)", BCP 195, RFC 7525, DOI 10.17487/RFC7525, May 2015, <<https://www.rfc-editor.org/info/rfc7525>>.
- [RFC8126] Cotton, M., Leiba, B., and T. Narten, "Guidelines for Writing an IANA Considerations Section in RFCs", BCP 26, RFC 8126, DOI 10.17487/RFC8126, June 2017, <<https://www.rfc-editor.org/info/rfc8126>>.

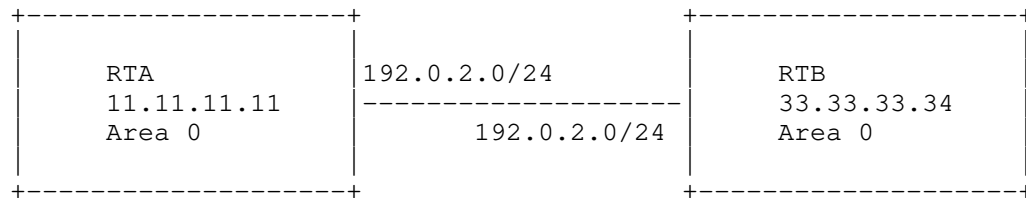
- [RFC8202] Ginsberg, L., Previdi, S., and W. Henderickx, "IS-IS Multi-Instance", RFC 8202, DOI 10.17487/RFC8202, June 2017, <<https://www.rfc-editor.org/info/rfc8202>>.
- [RFC8231] Crabbe, E., Minei, I., Medved, J., and R. Varga, "Path Computation Element Communication Protocol (PCEP) Extensions for Stateful PCE", RFC 8231, DOI 10.17487/RFC8231, September 2017, <<https://www.rfc-editor.org/info/rfc8231>>.
- [RFC8253] Lopez, D., Gonzalez de Dios, O., Wu, Q., and D. Dhody, "PCEPS: Usage of TLS to Provide a Secure Transport for the Path Computation Element Communication Protocol (PCEP)", RFC 8253, DOI 10.17487/RFC8253, October 2017, <<https://www.rfc-editor.org/info/rfc8253>>.
- [RFC8281] Crabbe, E., Minei, I., Sivabalan, S., and R. Varga, "Path Computation Element Communication Protocol (PCEP) Extensions for PCE-Initiated LSP Setup in a Stateful PCE Model", RFC 8281, DOI 10.17487/RFC8281, December 2017, <<https://www.rfc-editor.org/info/rfc8281>>.
- [RFC8283] Farrel, A., Ed., Zhao, Q., Ed., Li, Z., and C. Zhou, "An Architecture for Use of PCE and the PCE Communication Protocol (PCEP) in a Network with Central Control", RFC 8283, DOI 10.17487/RFC8283, December 2017, <<https://www.rfc-editor.org/info/rfc8283>>.
- [RFC8453] Ceccarelli, D., Ed. and Y. Lee, Ed., "Framework for Abstraction and Control of TE Networks (ACTN)", RFC 8453, DOI 10.17487/RFC8453, August 2018, <<https://www.rfc-editor.org/info/rfc8453>>.
- [RFC8637] Dhody, D., Lee, Y., and D. Ceccarelli, "Applicability of the Path Computation Element (PCE) to the Abstraction and Control of TE Networks (ACTN)", RFC 8637, DOI 10.17487/RFC8637, July 2019, <<https://www.rfc-editor.org/info/rfc8637>>.

Appendix A. Examples

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

A.1. All Nodes

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



RTA

LS Node

TLV - Local Node Descriptors

Sub-TLV - 514: OSPF Area-ID: 0.0.0.0

Sub-TLV - 515: IGP Router-ID: 11.11.11.11

TLV - Node Attributes TLV

Sub-TLV(s)

LS Link

TLV - Local Node Descriptors

Sub-TLV - 514: OSPF Area-ID: 0.0.0.0

Sub-TLV - 515: IGP Router-ID: 11.11.11.11

TLV - Remote Node Descriptors

Sub-TLV - 514: OSPF Area-ID: 0.0.0.0

Sub-TLV - 515: IGP Router-ID: 22.22.22.22

TLV - Link Descriptors

Sub-TLV - 259: IPv4 interface: 192.0.2.1

Sub-TLV - 260: IPv4 neighbor: 192.0.2.2

TLV - Link Attributes TLV

Sub-TLV(s)

RTB

LS Node

TLV - Local Node Descriptors

Sub-TLV - 514: OSPF Area-ID: 0.0.0.0

Sub-TLV - 515: IGP Router-ID: 22.22.22.22

TLV - Node Attributes TLV

Sub-TLV(s)

LS Link

TLV - Local Node Descriptors

Sub-TLV - 514: OSPF Area-ID: 0.0.0.0

```

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

```

A.2. Designated Node

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

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

A.3. Between PCEs

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

```

PCE1
----
LS Node
  TLV - Local Node Descriptors
    Sub-TLV - 512: Autonomous System: 100 (Domain 1)
    Sub-TLV - 515: IGP Router-ID: 11.11.11.11 (abstract)

LS Link
  TLV - Local Node Descriptors
    Sub-TLV - 512: Autonomous System: 100
    Sub-TLV - 515: IGP Router-ID: 11.11.11.11 (abstract)
  TLV - Remote Node Descriptors
    Sub-TLV - 512: Autonomous System: 200 (Domain 2)
    Sub-TLV - 515: IGP Router-ID: 22.22.22.22 (abstract)
  TLV - Link Descriptors
    Sub-TLV - 259: IPv4 interface: 192.0.2.1
    Sub-TLV - 260: IPv4 neighbor: 192.0.2.2
  TLV - Link Attributes TLV
    Sub-TLV(s)

```

LS Link

- TLV - Local Node Descriptors
 - Sub-TLV - 512: Autonomous System: 100
 - Sub-TLV - 515: IGP Router-ID: 11.11.11.11 (abstract)
- TLV - Remote Node Descriptors
 - Sub-TLV - 512: Autonomous System: 200
 - Sub-TLV - 515: IGP Router-ID: 22.22.22.22 (abstract)
- TLV - Link Descriptors
 - Sub-TLV - 259: IPv4 interface: 198.51.100.1
 - Sub-TLV - 260: IPv4 neighbor: 198.51.100.2
- TLV - Link Attributes TLV
 - Sub-TLV(s)

LS Link

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

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

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

Appendix B. Contributor Addresses

Udayasree Palle

Email: udayasreereddy@gmail.com

Sergio Belotti
Nokia

Email: sergio.belotti@nokia.com

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

Email: satishk@huawei.com

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

Email: c.l@huawei.com

Authors' Addresses

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

Email: dhruv.ietf@gmail.com

Shuping Peng
Huawei Technologies
Huawei Bld., No.156 Beiqing Rd.
Beijing
100095
China

Email: pengshuping@huawei.com

Young Lee
Samsung Electronics
Seoul
South Korea

Email: younglee.tx@gmail.com

Daniele Ceccarelli
Ericsson
Torshamnsgatan, 48
Stockholm
Sweden

Email: daniele.ceccarelli@ericsson.com

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

Email: wangaijun@tsinghua.org.cn

Gyan Mishra
Verizon Inc.

Email: gyan.s.mishra@verizon.com

Siva Sivabalan
Ciena Corporation

Email: ssivabal@ciena.com

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

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

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

Abstract

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

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

Requirements Language

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

Status of This Memo

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

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

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

This Internet-Draft will expire on 6 September 2022.

Copyright Notice

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

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

Table of Contents

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

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

1. Introduction

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

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

Section 4.3 of [RFC4655] describes the potential load of the TED on a network node and proposes an architecture where the TED is maintained by the PCE rather than the network nodes. However, it does not describe how a PCE would obtain the information needed to populate its TED. PCE may construct its TED by participating in the IGP ([RFC3630] and [RFC5305] for MPLS-TE; [RFC4203] and [RFC5307] for GMPLS). An alternative mechanism is offered by BGP-LS [I-D.ietf-idr-rfc7752bis] .

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

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

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

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

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

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

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

1.1. Scope

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

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

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

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

2. Terminology

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

3. Applicability

The mechanism specified in this draft is applicable to deployments:

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

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

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

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

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

4. Requirements for PCEP extensions

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

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

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

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

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

6. Overview of Extensions to PCEP

6.1. New Messages

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

6.2. Capability Advertisement

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

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

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

6.3. Initial Link-State (and TE) Synchronization

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

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

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

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

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

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

The successful LS Synchronization sequence is shown in Figure 1.

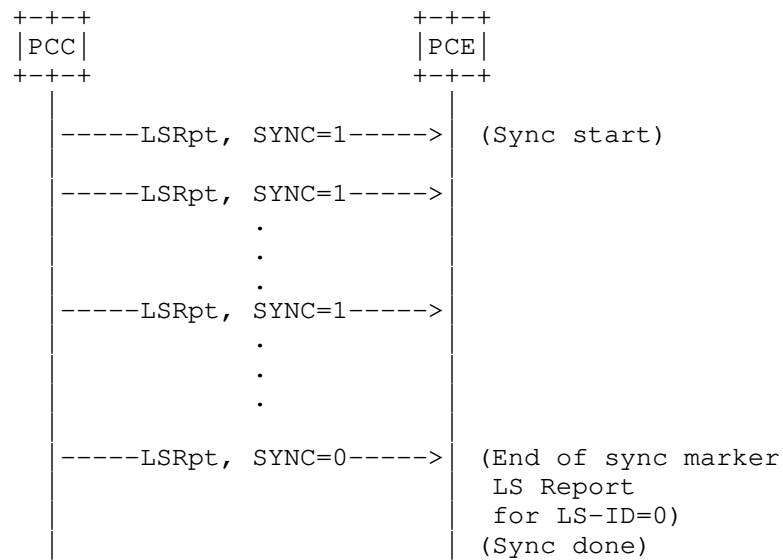


Figure 1: Successful LS synchronization

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

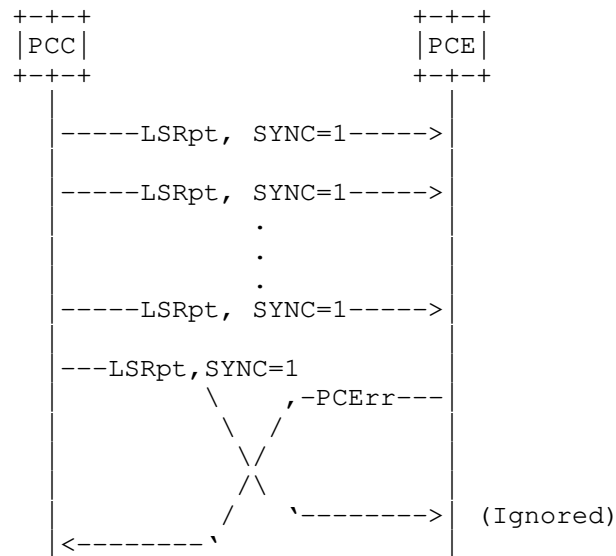


Figure 2: Failed LS synchronization (PCE failure)

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

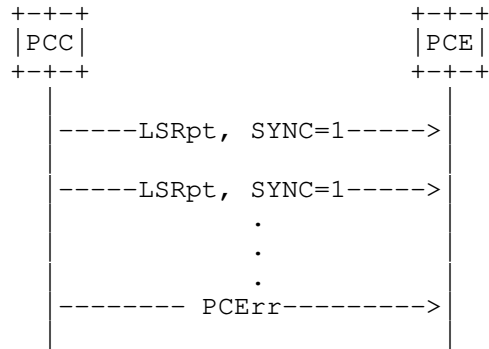


Figure 3: Failed LS synchronization (PCC failure)

6.3.1. Optimizations for LS Synchronization

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

6.4. LS Report

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

More details about the LSRpt message are in Section 8.1.

7. Transport

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

8. PCEP Messages

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

8.1. LS Report Message

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

The format of the LSRpt message is as follows:

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

Where:

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

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

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

8.2. The PCErr Message

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

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

```

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

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

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

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

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

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

```

9. Objects and TLV

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

9.1. TLV Format

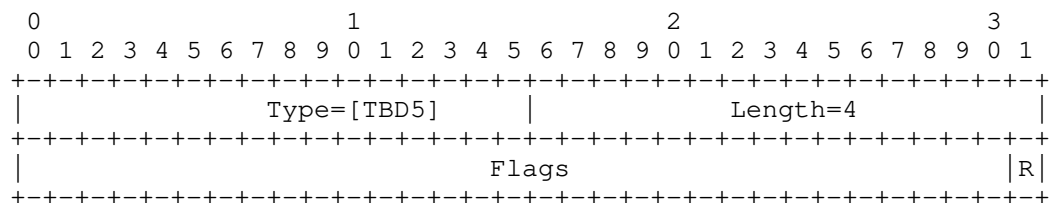
The TLV and the sub-TLV format (and padding) in this document, is as per section 7.1 of [RFC5440].

9.2. Open Object

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

9.2.1. LS Capability TLV

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



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

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

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

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

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

9.3. LS Object

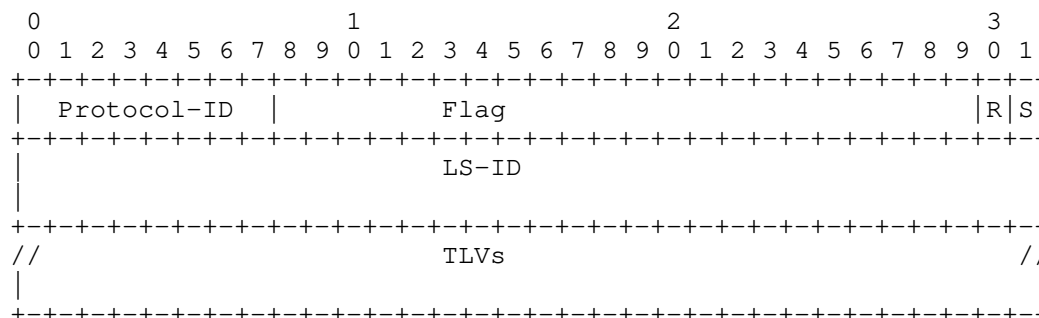
The LS (link-state) object MUST be carried within LSRpt messages and MAY be carried within PCErr messages. The LS object contains a set of fields used to specify the target node or link. It also contains a flag indicating to a PCE that the LS synchronization is in progress. The TLVs used with the LS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

LS Object-Class is TBD6.

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

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

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



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

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

Flags (24-bit):

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

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

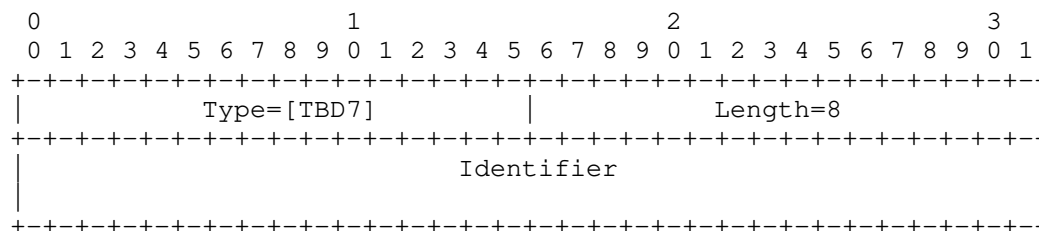
Unassigned bits are considered reserved. They MUST be set to 0 on transmission and MUST be ignored on receipt.

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

9.3.1. Routing Universe TLV

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

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



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

Identifier	Routing Universe
0	Default Layer 3 Routing topology

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

9.3.2. Route Distinguisher TLV

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

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

9.3.3. Virtual Network TLV

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

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

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

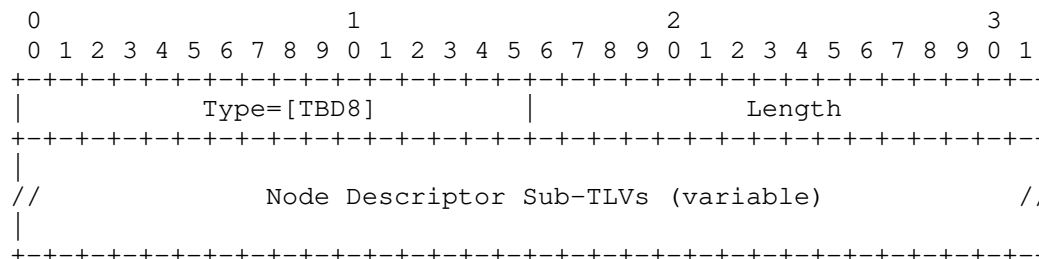
9.3.4. Local Node Descriptors TLV

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

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

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

This TLV is applicable for all LS Object-Type.

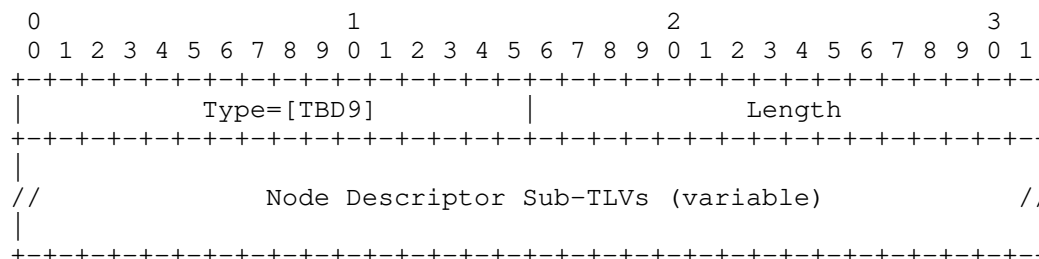


The value contains one or more Node Descriptor Sub-TLVs defined in Section 9.3.6.

9.3.5. Remote Node Descriptors TLV

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

This TLV is applicable for LS Link Object-Type.



9.3.6. Node Descriptors Sub-TLVs

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

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

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

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

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

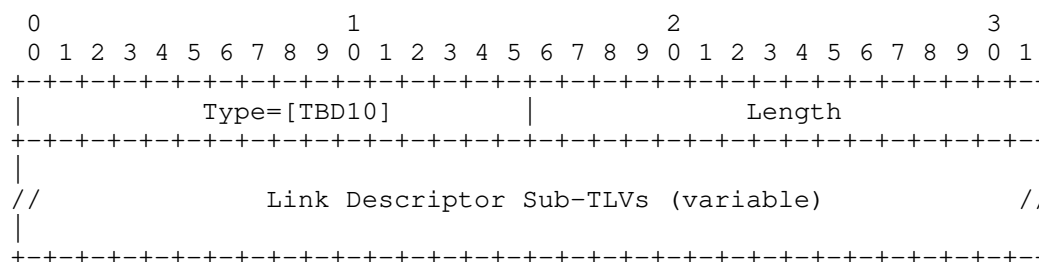
A new sub-TLV type (1) is allocated for SPEAKER-ENTITY-ID sub-TLV. The length and value fields are as per [RFC8232].

9.3.7. Link Descriptors TLV

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

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

This TLV is applicable for LS Link Object-Type.



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

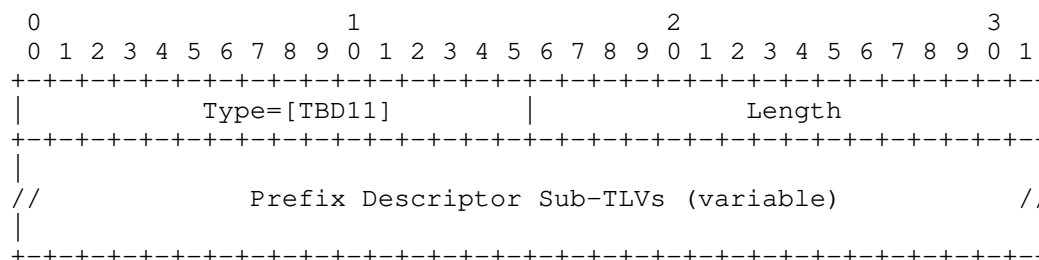
The format and semantics of the 'value' fields in most 'Link Descriptor' sub-TLVs correspond to the format and semantics of value fields in IS-IS Extended IS Reachability sub-TLVs, defined in [RFC5305], [RFC5307] and [RFC6119]. Although the encodings for 'Link Descriptor' TLVs were originally defined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

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

9.3.8. Prefix Descriptors TLV

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

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

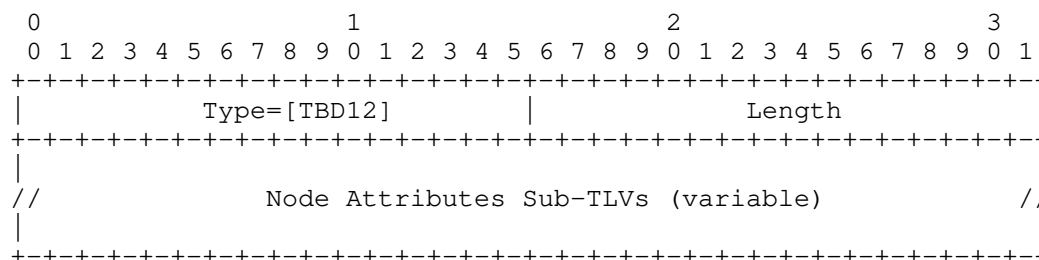


All Prefix Descriptors TLVs defined for BGP-LS can then be used with PCEP-LS as well. No new PCEP sub-TLVs for Prefix Descriptor are defined in this document.

9.3.9. PCEP-LS Attributes

9.3.9.1. Node Attributes TLV

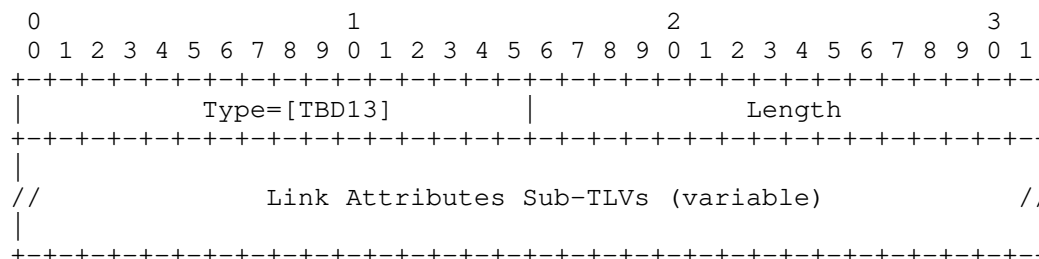
This is an optional attribute that is used to carry node attributes. This TLV is applicable for LS Node Object-Type.



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

9.3.9.2. Link Attributes TLV

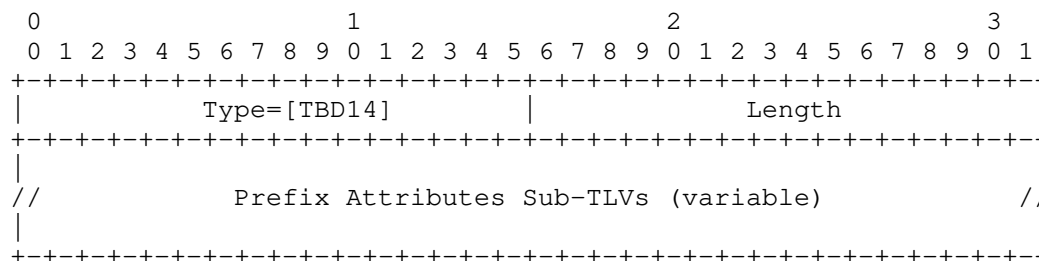
This TLV is applicable for LS Link Object-Type. The format and semantics of the 'value' fields in some 'Link Attribute' sub-TLVs correspond to the format and semantics of the 'value' fields in IS-IS Extended IS Reachability sub-TLVs, defined in [RFC5305], [RFC5307] and [I-D.ietf-idr-rfc7752bis]. Although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



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

9.3.9.3. Prefix Attributes TLV

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



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

9.3.10. Removal of an Attribute

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

10. Other Considerations

10.1. Inter-AS Links

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

11. Security Considerations

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

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

12. Manageability Considerations

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

12.1. Control of Function and Policy

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

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

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

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

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

12.2. Information and Data Models

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

An implementation SHOULD also provide the statistics:

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

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

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

12.3. Liveness Detection and Monitoring

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

12.4. Verify Correct Operations

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

12.5. Requirements On Other Protocols

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

12.6. Impact On Network Operations

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

13. IANA Considerations

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

13.1. PCEP Messages

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

Value	Meaning	Reference
TBD3	LSRpt	[This I-D]

13.2. PCEP Objects

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

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

13.3. LS Object

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

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

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

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

The following values are defined in this document:

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

13.4. PCEP-Error Object

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

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

13.5. PCEP TLV Type Indicators

This document defines the following new PCEP TLVs.

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

13.6. PCEP-LS Sub-TLV Type Indicators

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

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

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

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

Value	Meaning
1	SPEAKER-ENTITY-ID

14. TLV Code Points Summary

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

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

* this TLV is defined in a different PCEP document

Figure 4: TLV Table

15. Implementation Status

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

15.1. Hierarchical Transport PCE controllers

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

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

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

15.2. ONOS-based Controller (MDSC and PNC)

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

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

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

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

16. Acknowledgments

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

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

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

17. References

17.1. Normative References

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

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <<https://www.rfc-editor.org/info/rfc2119>>.
- [RFC5305] Li, T. and H. Smit, "IS-IS Extensions for Traffic Engineering", RFC 5305, DOI 10.17487/RFC5305, October 2008, <<https://www.rfc-editor.org/info/rfc5305>>.
- [RFC5307] Kompella, K., Ed. and Y. Rekhter, Ed., "IS-IS Extensions in Support of Generalized Multi-Protocol Label Switching (GMPLS)", RFC 5307, DOI 10.17487/RFC5307, October 2008, <<https://www.rfc-editor.org/info/rfc5307>>.
- [RFC5440] Vasseur, JP., Ed. and JL. Le Roux, Ed., "Path Computation Element (PCE) Communication Protocol (PCEP)", RFC 5440, DOI 10.17487/RFC5440, March 2009, <<https://www.rfc-editor.org/info/rfc5440>>.
- [RFC6119] Harrison, J., Berger, J., and M. Bartlett, "IPv6 Traffic Engineering in IS-IS", RFC 6119, DOI 10.17487/RFC6119, February 2011, <<https://www.rfc-editor.org/info/rfc6119>>.
- [RFC8174] Leiba, B., "Ambiguity of Uppercase vs Lowercase in RFC 2119 Key Words", BCP 14, RFC 8174, DOI 10.17487/RFC8174, May 2017, <<https://www.rfc-editor.org/info/rfc8174>>.
- [RFC8232] Crabbe, E., Minei, I., Medved, J., Varga, R., Zhang, X., and D. Dhody, "Optimizations of Label Switched Path State Synchronization Procedures for a Stateful PCE", RFC 8232, DOI 10.17487/RFC8232, September 2017, <<https://www.rfc-editor.org/info/rfc8232>>.

17.2. Informative References

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

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

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

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

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

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

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

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

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

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

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

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

- [RFC5925] Touch, J., Mankin, A., and R. Bonica, "The TCP Authentication Option", RFC 5925, DOI 10.17487/RFC5925, June 2010, <<https://www.rfc-editor.org/info/rfc5925>>.
- [RFC6549] Lindem, A., Roy, A., and S. Mirtorabi, "OSPFv2 Multi-Instance Extensions", RFC 6549, DOI 10.17487/RFC6549, March 2012, <<https://www.rfc-editor.org/info/rfc6549>>.
- [RFC6805] King, D., Ed. and A. Farrel, Ed., "The Application of the Path Computation Element Architecture to the Determination of a Sequence of Domains in MPLS and GMPLS", RFC 6805, DOI 10.17487/RFC6805, November 2012, <<https://www.rfc-editor.org/info/rfc6805>>.
- [RFC6952] Jethanandani, M., Patel, K., and L. Zheng, "Analysis of BGP, LDP, PCEP, and MSDP Issues According to the Keying and Authentication for Routing Protocols (KARP) Design Guide", RFC 6952, DOI 10.17487/RFC6952, May 2013, <<https://www.rfc-editor.org/info/rfc6952>>.
- [RFC7525] Sheffer, Y., Holz, R., and P. Saint-Andre, "Recommendations for Secure Use of Transport Layer Security (TLS) and Datagram Transport Layer Security (DTLS)", BCP 195, RFC 7525, DOI 10.17487/RFC7525, May 2015, <<https://www.rfc-editor.org/info/rfc7525>>.
- [RFC8126] Cotton, M., Leiba, B., and T. Narten, "Guidelines for Writing an IANA Considerations Section in RFCs", BCP 26, RFC 8126, DOI 10.17487/RFC8126, June 2017, <<https://www.rfc-editor.org/info/rfc8126>>.
- [RFC8202] Ginsberg, L., Previdi, S., and W. Henderickx, "IS-IS Multi-Instance", RFC 8202, DOI 10.17487/RFC8202, June 2017, <<https://www.rfc-editor.org/info/rfc8202>>.
- [RFC8231] Crabbe, E., Minei, I., Medved, J., and R. Varga, "Path Computation Element Communication Protocol (PCEP) Extensions for Stateful PCE", RFC 8231, DOI 10.17487/RFC8231, September 2017, <<https://www.rfc-editor.org/info/rfc8231>>.
- [RFC8253] Lopez, D., Gonzalez de Dios, O., Wu, Q., and D. Dhody, "PCEPS: Usage of TLS to Provide a Secure Transport for the Path Computation Element Communication Protocol (PCEP)", RFC 8253, DOI 10.17487/RFC8253, October 2017, <<https://www.rfc-editor.org/info/rfc8253>>.

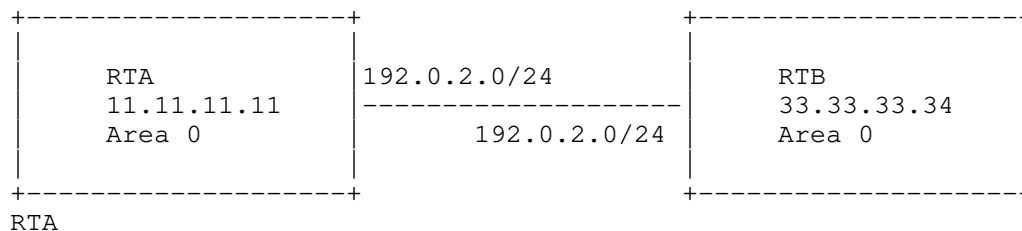
- [RFC8281] Crabbe, E., Minei, I., Sivabalan, S., and R. Varga, "Path Computation Element Communication Protocol (PCEP) Extensions for PCE-Initiated LSP Setup in a Stateful PCE Model", RFC 8281, DOI 10.17487/RFC8281, December 2017, <<https://www.rfc-editor.org/info/rfc8281>>.
- [RFC8283] Farrel, A.,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

Appendix A. Examples

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

A.1. All Nodes

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



LS Node

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

LS Link

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

RTB

LS Node

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

LS Link

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

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

A.2. Designated Node

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

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

A.3. Between PCEs

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

PCE1

LS Node

TLV - Local Node Descriptors

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

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

LS Link

TLV - Local Node Descriptors

Sub-TLV - 512: Autonomous System: 100

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

TLV - Remote Node Descriptors

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

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

TLV - Link Descriptors

Sub-TLV - 259: IPv4 interface: 192.0.2.1

Sub-TLV - 260: IPv4 neighbor: 192.0.2.2

TLV - Link Attributes TLV

Sub-TLV(s)

LS Link

TLV - Local Node Descriptors

Sub-TLV - 512: Autonomous System: 100

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

TLV - Remote Node Descriptors

Sub-TLV - 512: Autonomous System: 200

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

TLV - Link Descriptors

Sub-TLV - 259: IPv4 interface: 198.51.100.1

Sub-TLV - 260: IPv4 neighbor: 198.51.100.2

TLV - Link Attributes TLV
Sub-TLV(s)

LS Link

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

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

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

Appendix B. Contributor Addresses

Udayasree Palle

Email: udayasreereddy@gmail.com

Sergio Belotti
Nokia

Email: sergio.belotti@nokia.com

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

Email: satishk@huawei.com

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

Email: c.l@huawei.com

Authors' Addresses

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

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

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

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

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

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

Siva Sivabalan
Ciena Corporation
Email: ssivabal@ciena.com

PCE Working Group
Internet-Draft
Intended status: Standards Track
Expires: April 26, 2022

J. Dong
S. Fang
Huawei Technologies
L. Han
M. Wang
China Mobile
October 23, 2021

Support for Virtual Transport Network (VTN) in the Path Computation
Element Communication Protocol (PCEP)
draft-dong-pce-pcep-vtn-00

Abstract

With the introduction and evolvement of 5G and other network scenarios, some existing or new customers may require connectivity services with advanced characteristics comparing to traditional Virtual Private Networks (VPNs). Such kind of network service is called enhanced VPNs (VPN+). The typical application of VPN+ is to provide network slice services.

A Virtual Transport Network (VTN) is a virtual underlay network which consists of a set of dedicated or shared network resources allocated from the physical underlay network, and is associated with a customized logical network topology. VPN+ services can be delivered by mapping one or a group of overlay VPNs to the appropriate VTNs as the virtual underlay. Then traffic flows of the VPN+ service can be steered onto the TE paths within the VTN.

The Path Computation Element (PCE) provides path computation functions in support of traffic engineering in Multiprotocol Label Switching (MPLS), Generalized MPLS (GMPLS) and Segment Routing (SR) networks.

This document specifies the extensions to PCE communication Protocol (PCEP) to carry VTN information in the PCEP messages. The extensions in this document can be used in the basic PCE computation, the stateful PCE and the PCE-initiated LSP mechanisms to indicate path computation, path status report and path initialization within a specific VTN.

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 April 26, 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	4
2.1. New TLV in LSPA Object	4
2.2. Capability Advertisement	5
3. Operations	6
4. Security Considerations	6
5. IANA Considerations	7
6. Contributors	7
7. Acknowledgments	7
8. References	7
8.1. Normative References	7

8.2. Informative References	8
Authors' Addresses	9

1. Introduction

[RFC5440] describes the Path Computation Element (PCE) Communication Protocol (PCEP). PCEP enables the communication between a Path Computation Client (PCC) and a PCE, or between PCE and PCE, for the purpose of computation of Multiprotocol Label Switching (MPLS) as well as Generalized MPLS (GMPLS) Traffic Engineering Label Switched Path (TE LSP) characteristics.

[RFC8231] specifies a set of extensions to PCEP to enable stateful control of TE LSPs within and across PCEP sessions in compliance with [RFC4657]. It includes mechanisms to effect LSP State Synchronization between PCCs and PCEs, delegation of control over LSPs to PCEs, and PCE control of timing and sequence of path computations within and across PCEP sessions. The model of operation where LSPs are initiated from the PCE is described in [RFC8281]. [RFC8664] specifies PCEP extensions to allow a stateful PCE to compute and initiate TE paths, as well as a PCC to request a path subject to certain constraints and optimization criteria in SR networks.

With the introduction and evolvement of 5G and other network scenarios, some existing or new customers may require connectivity services with advanced characteristics comparing to traditional Virtual Private Networks (VPNs). Such kind of network service is called enhanced VPNs (VPN+). The typical application of VPN+ is to provide network slice services. The concept and general framework of IETF network slice are described in [I-D.ietf-teas-ietf-network-slices].

[I-D.ietf-teas-enhanced-vpn] describes a framework and the candidate component technologies for providing VPN+ services. It also introduces the concept of Virtual Transport Network (VTN). A Virtual Transport Network (VTN) is a virtual underlay network which consists of a set of dedicated or shared network resources allocated from the physical underlay network, and is associated with a customized logical network topology. VPN+ services can be delivered by mapping one or a group of overlay VPNs to the appropriate VTNs as the underlay, so as to provide the network characteristics required by the customers. Then the traffic flows of the VPN+ service can be steered onto the TE paths within the VTN.

In MPLS or SR based network, the set of network resources allocated to a VTN can be identified using resource-aware SR SIDs as defined in [I-D.ietf-spring-resource-aware-segments]

[I-D.ietf-spring-sr-for-enhanced-vpn], or the VTN Resource ID as defined in [I-D.dong-6man-enhanced-vpn-vtn-id]. The logical topology associated with a VTN could be specified using mechanisms such as Multi-Topology [RFC4915], [RFC5120] or Flex-Algo [I-D.ietf-lsr-flex-algo], etc.

To meet specific service requirement, traffic flows of a VPN+ service need be steered onto TE paths of the corresponding VTN. A PCC may request the PCE for computing a TE path within a VTN, so that the path computation would take the resource attribute and the associated topology of the VTN into consideration. Correspondingly, a PCE may reply or initiate a TE path with VTN-specific control and data plane information to a PCC.

This document extends PCEP to allow VTN information to be encoded in the PCEP messages. The extensions in this document can be used in the basic PCE computation, the stateful PCE and the PCE-initiated LSP mechanisms to indicate path computation, path status report and path initialization within the context of a specific VTN.

2. PCEP Extensions

2.1. New TLV in LSPA Object

A new VTN TLV for use in the LSPA Object is defined to indicate the VTN ID and the related information as constraints. The format of the VTN TLV is as follows:

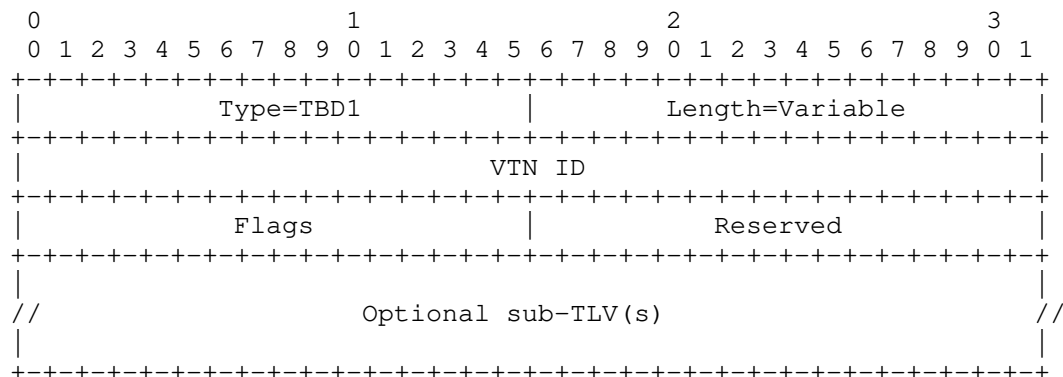


Figure 1: VTN TLV Format

Where:

- o VTN ID: A global significant 32-bit identifier which is used to identify a VTN.

- o **Flags:** 16-bit flags. Currently all the flags are reserved for future use. They SHOULD be set to zero on transmission and MUST be ignored on receipt.
- o **Reserved:** 16-bit reserved field for future use. All the bits SHOULD be set to zero on transmission and MUST be ignored on receipt.
- o **Optional sub-TLVs:** Additional information which can be used in as VTN-specific constraints. Currently no sub-TLV is defined in this document.

2.2. Capability Advertisement

A PCEP speaker indicates whether it supports VTN specific path computation using a new PCEP capability called "VTN-CAPABILITY". When the PCEP session is created, it sends an Open message with an OPEN Object containing the VTN-CAPABILITY TLV. The format of this TLV is as follows:

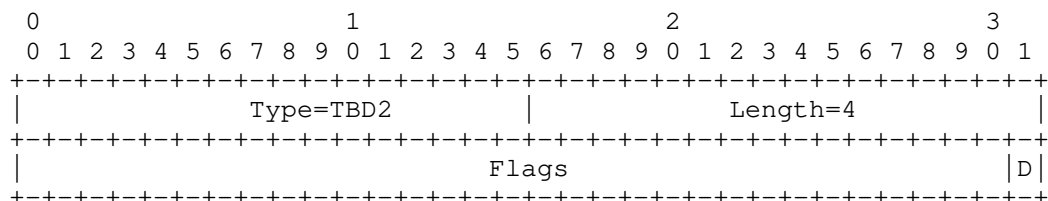


Figure 2: VTN CAPABILITY TLV

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

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

- o **D (Data Plane VTN-ID CAPABILITY - 1 bit):** if set to 1 by a PCC, the D flag indicates that the PCC supports the encapsulation of data plane VTN-ID in data packet; if set to 1 by a PCE, the D flag indicates that the PCE supports to provide path computation result with the data plane VTN-ID.
- o **Unassigned bits in the Flags field MUST be set to zero and ignored on receipt.**

3. Operations

The new VTN TLV defined in this document can be used in the basic PCE computation, the stateful PCE and the PCE-initiated LSP mechanisms to indicate path computation, path status report and path initialization within the context of a specific VTN.

Information about the VTN-specific network resource and topology attributes can be obtained by the PCE either from the network planning system, or using a distributed control plane such as IGP or BGP-LS with necessary extensions. The detailed mechanism is out of the scope of this document. The obtained VTN specific attributes can be used in path computation when the VTN-ID is specified in the path computation request.

With the basic path computation mechanism, the new VTN TLV can be used to indicate the VTN in which the path computation is requested. In a PCReq message, the VTN TLV MAY be carried in the LSPA Object to indicate that the path computation needs to be executed using the resource and topological attributes of the VTN. The PCE SHOULD use the network resource and topology attributes associated with the specified VTN as the parameters of path computation. In a PCRep message, the VTN TLV MAY be carried in the LSPA Object in case of failure to indicate the path computation in the VTN was not successful.

The new VTN TLV can also be used in the stateful PCE mechanisms. A PCC MAY include the VTN TLV in PCRpt message to indicate the VTN in which the TE path is reported. And A PCE MAY include the VTN TLV in PCUpd Message to indicate the VTN in which the TE path needs to be updated.

With the PCE-Initiated LSP mechanism, the PCE MAY include the VTN TLV in PCInitiate or PCUpd message to indicate the VTN in which the path is computed, so that the PCC will use the VTN-specific resources and data plane VTN-ID in constructing or updating the TE path.

4. Security Considerations

This document defines a new VTN TLV that do not add any new security concerns beyond those discussed in [RFC5440] in itself. Some deployments may find the VTN information to be extra sensitive and could be used to influence path computation and setup with adverse effect. Additionally, snooping of PCEP messages with such data or using PCEP messages for network reconnaissance may give an attacker sensitive information about the operations of the network. Thus, such deployment should employ suitable PCEP security mechanisms like TCP Authentication Option (TCP-AO) [RFC5925] or Transport Layer

Security (TLS) [RFC8253]. The procedure based on TLS is considered a security enhancement and thus is much better suited for the sensitive information.

5. IANA Considerations

This document makes following requests to IANA for action.

IANA is requested to make the following allocations in the "PCEP TLV Type Indicators" subregistry of the "Path Computation Element Protocol (PCEP) Numbers" registry:

Value	Description	Reference
-----	-----	-----
TBD1	VTN	This document
TBD2	VTN CAPABILITY	This document

6. Contributors

Dhruv Dhody
Email: dhruv.ietf@gmail.com

Zhibo Hu
Email: huzhibo@huawei.com

7. Acknowledgments

The authors would like to thank Zhenbin Li for his review and valuable comments.

8. References

8.1. Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <<https://www.rfc-editor.org/info/rfc2119>>.
- [RFC5440] Vasseur, JP., Ed. and JL. Le Roux, Ed., "Path Computation Element (PCE) Communication Protocol (PCEP)", RFC 5440, DOI 10.17487/RFC5440, March 2009, <<https://www.rfc-editor.org/info/rfc5440>>.
- [RFC8174] Leiba, B., "Ambiguity of Uppercase vs Lowercase in RFC 2119 Key Words", BCP 14, RFC 8174, DOI 10.17487/RFC8174, May 2017, <<https://www.rfc-editor.org/info/rfc8174>>.

- [RFC8231] Crabbe, E., Minei, I., Medved, J., and R. Varga, "Path Computation Element Communication Protocol (PCEP) Extensions for Stateful PCE", RFC 8231, DOI 10.17487/RFC8231, September 2017, <<https://www.rfc-editor.org/info/rfc8231>>.
- [RFC8281] Crabbe, E., Minei, I., Sivabalan, S., and R. Varga, "Path Computation Element Communication Protocol (PCEP) Extensions for PCE-Initiated LSP Setup in a Stateful PCE Model", RFC 8281, DOI 10.17487/RFC8281, December 2017, <<https://www.rfc-editor.org/info/rfc8281>>.
- [RFC8664] Sivabalan, S., Filsfils, C., Tantsura, J., Henderickx, W., and J. Hardwick, "Path Computation Element Communication Protocol (PCEP) Extensions for Segment Routing", RFC 8664, DOI 10.17487/RFC8664, December 2019, <<https://www.rfc-editor.org/info/rfc8664>>.
- [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-17 (work in progress), July 2021.

8.2. Informative References

- [RFC4657] Ash, J., Ed. and J. Le Roux, Ed., "Path Computation Element (PCE) Communication Protocol Generic Requirements", RFC 4657, DOI 10.17487/RFC4657, September 2006, <<https://www.rfc-editor.org/info/rfc4657>>.
- [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-ISs)", RFC 5120, DOI 10.17487/RFC5120, February 2008, <<https://www.rfc-editor.org/info/rfc5120>>.
- [RFC5925] Touch, J., Mankin, A., and R. Bonica, "The TCP Authentication Option", RFC 5925, DOI 10.17487/RFC5925, June 2010, <<https://www.rfc-editor.org/info/rfc5925>>.

- [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>>.
- [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>>.
- [I-D.ietf-teas-ietf-network-slices]
Farrel, A., Gray, E., Drake, J., Rokui, R., Homma, S., Makhiyani, K., Contreras, L. M., and J. Tantsura, "Framework for IETF Network Slices", draft-ietf-teas-ietf-network-slices-04 (work in progress), August 2021.
- [I-D.ietf-teas-enhanced-vpn]
Dong, J., Bryant, S., Li, Z., Miyasaka, T., and Y. Lee, "A Framework for Enhanced Virtual Private Network (VPN+) Services", draft-ietf-teas-enhanced-vpn-08 (work in progress), July 2021.
- [I-D.ietf-spring-resource-aware-segments]
Dong, J., Bryant, S., Miyasaka, T., Zhu, Y., Qin, F., Li, Z., and F. Clad, "Introducing Resource Awareness to SR Segments", draft-ietf-spring-resource-aware-segments-03 (work in progress), July 2021.
- [I-D.ietf-spring-sr-for-enhanced-vpn]
Dong, J., Bryant, S., Miyasaka, T., Zhu, Y., Qin, F., Li, Z., and F. Clad, "Segment Routing based Virtual Transport Network (VTN) for Enhanced VPN", draft-ietf-spring-sr-for-enhanced-vpn-01 (work in progress), July 2021.
- [I-D.dong-6man-enhanced-vpn-vtn-id]
Dong, J., Li, Z., Xie, C., Ma, C., and G. Mishra, "Carrying Virtual Transport Network Identifier in IPv6 Extension Header", draft-dong-6man-enhanced-vpn-vtn-id-05 (work in progress), September 2021.

Authors' Addresses

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

Email: jie.dong@huawei.com

Sheng Fang
Huawei Technologies
Huawei Campus, No. 156 Beiqing Rd.
Beijing 100095
China

Email: fangsheng@huawei.com

Liuyan Han
China Mobile
Beijing
China

Email: hanliuyan@chinamobile.com

Minxue Wang
China Mobile
Beijing
China

Email: wangminxue@chinamobile.com

PCE Working Group
Internet-Draft
Intended status: Standards Track
Expires: 28 April 2022

M. Koldychev
Cisco Systems, Inc.
S. Sivabalan
Ciena Corporation
T. Saad
V. Beeram
Juniper Networks, Inc.
H. Bidgoli
Nokia
B. Yadav
Ciena
S. Peng
Huawei Technologies
25 October 2021

PCEP Extensions for Signaling Multipath Information
draft-ietf-pce-multipath-03

Abstract

Path computation algorithms are not limited to return a single optimal path. Multiple paths may exist that satisfy the given objectives and constraints. This document defines a mechanism to encode multiple paths for a single set of objectives and constraints. This is a generic PCEP mechanism, not specific to any path setup type or dataplane. The mechanism is applicable to both stateless and stateful PCEP.

Status of This Memo

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

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at <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 28 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	4
2.1. Terms and Abbreviations	4
3. Motivation	4
3.1. Signaling Multiple Segment-Lists of an SR Candidate-Path	4
3.2. Splitting of Requested Bandwidth	4
3.3. Providing Backup path for Protection	4
4. Protocol Extensions	5
4.1. Multipath Capability TLV	5
4.2. Path Attributes Object	6
4.3. Multipath Weight TLV	6
4.4. Multipath Backup TLV	7
4.5. Multipath Opposite Direction Path TLV	8
4.6. Composite Candidate Path	9
5. Operation	10
5.1. Signaling Multiple Paths for Loadbalancing	10
5.2. Signaling Multiple Paths for Protection	11
6. PCEP Message Extensions	12
7. Examples	12
7.1. SR Policy Candidate-Path with Multiple Segment-Lists	12
7.2. Two Primary Paths Protected by One Backup Path	13
7.3. Composite Candidate Path	14
7.4. Opposite Direction Tunnels	15
8. IANA Considerations	16
8.1. PCEP Object	17
8.2. PCEP TLV	17
8.3. PCEP-Error Object	17
8.4. Flags in the Multipath Capability TLV	18
8.5. Flags in the Path Attribute Object	18
8.6. Flags in the Multipath Backup TLV	18
8.7. Flags in the Multipath Opposite Direction Path TLV	19

9. Security Considerations	19
10. Acknowledgement	19
11. Contributors	19
12. References	19
12.1. Normative References	19
12.2. Informative References	21
Authors' Addresses	21

1. Introduction

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

PCEP Extensions for the Stateful PCE Model [RFC8231] describes a set of extensions to PCEP that enable active control of Multiprotocol Label Switching Traffic Engineering (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.

PCEP Extensions for Segment Routing [RFC8664] specifies extensions to the Path Computation Element Protocol (PCEP) that allow a stateful PCE to compute and initiate Traffic Engineering (TE) paths, as well as for a PCC to request a path subject to certain constraint(s) and optimization criteria in SR networks.

Segment Routing Policy for Traffic Engineering [I-D.ietf-spring-segment-routing-policy] details the concepts of SR Policy and approaches to steering traffic into an SR Policy. In particular, it describes the SR candidate-path as a collection of one or more Segment-Lists. The current PCEP standards only allow for signaling of one Segment-List per Candidate-Path. PCEP extension to support Segment Routing Policy Candidate Paths [I-D.ietf-pce-segment-routing-policy-cp] specifically avoids defining how to signal multipath information, and states that this will be defined in another document.

This document defines the required extensions that allow the signaling of multipath information via PCEP.

2. Terminology

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.

2.1. Terms and Abbreviations

The following terms are used in this document:

PCEP Tunnel:

The object identified by the PLSP-ID, see [I-D.koldychev-pce-operational] for more details.

3. Motivation

This extension is motivated by the use-cases described below.

3.1. Signaling Multiple Segment-Lists of an SR Candidate-Path

The Candidate-Path of an SR Policy is the unit of report/update in PCEP, see [I-D.ietf-pce-segment-routing-policy-cp]. Each Candidate-Path can contain multiple Segment-Lists and each Segment-List is encoded by one ERO. However, each PCEP LSP can contain only a single ERO, which prevents us from encoding multiple Segment-Lists within the same SR Candidate-Path.

With the help of the protocol extensions defined in this document, this limitation is overcome.

3.2. Splitting of Requested Bandwidth

A PCC may request a path with 80 Gbps of bandwidth, but all links in the network have only 50 Gbps capacity. The PCE can return two paths, that can together carry 80 Gbps. The PCC can then equally or unequally split the incoming 80 Gbps of traffic among the two paths. Section 4.3 introduces a new TLV that carries the path weight that allows for distribution of incoming traffic on to the multiple paths.

3.3. Providing Backup path for Protection

It is desirable for the PCE to compute and signal to the PCC a backup path that is used to protect a primary path within the multipaths in a given LSP.

Note that [RFC8745] specify the Path Protection association among LSPs. The use of [RFC8745] with multipath is out of scope of this document and is for future study.

When multipath is used, a backup path may protect one or more primary paths. For this reason, primary and backup path identifiers are needed to indicate which backup path(s) protect which primary path(s). Section 4.4 introduces a new TLV that carries the required information.

4. Protocol Extensions

4.1. Multipath Capability TLV

We define the MULTIPATH-CAP TLV that MAY be present in the OPEN object and/or the LSP object. The purpose of this TLV is two-fold:

1. From PCC: it tells how many multipaths per PCEP Tunnel, the PCC can install in forwarding.
2. From PCE: it tells that the PCE supports this standard and how many multipaths per PCEP Tunnel, the PCE can compute.

Only the first instance of this TLV can be processed, subsequent instances SHOULD be ignored.

Section 5 specify the usage of this TLV with Open message (within the OPEN object) and other PCEP messages (within the LSP object).

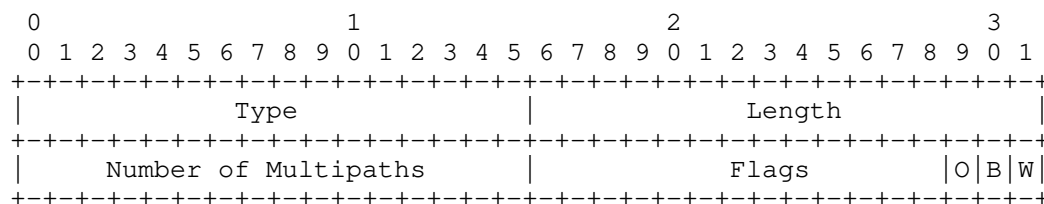


Figure 1: MULTIPATH-CAP TLV format

Type: TBD1 for "MULTIPATH-CAP" TLV.

Length: 4.

Number of Multipaths: the maximum number of multipaths per PCEP Tunnel. The value 0 indicates unlimited number.

W-flag: whether MULTIPATH-WEIGHT-TLV is supported.

B-flag: whether MULTIPATH-BACKUP-TLV is supported.

O-flag: whether MULTIPATH-OPPPDIR-PATH-TLV is supported.

4.2. Path Attributes Object

We define the PATH-ATTRIB object that is used to carry per-path information and to act as a separator between several ERO/RRO objects in the <intended-path>/<actual-path> RBNF element. The PATH-ATTRIB object always precedes the ERO/RRO that it applies to. If multiple ERO/RRO objects are present, then each ERO/RRO object MUST be preceded by an PATH-ATTRIB object that describes it.

The PATH-ATTRIB Object-Class value is TBD2.

The PATH-ATTRIB Object-Type value is 1.

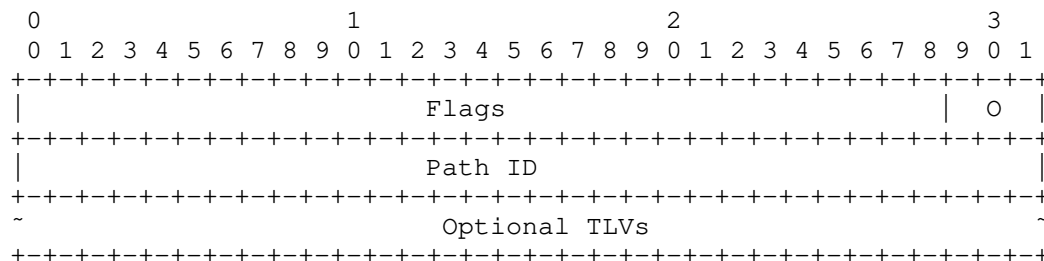


Figure 2: PATH-ATTRIB object format

O (Operational - 3 bits): operational state of the path, same values as the identically named field in the LSP object [RFC8231].

Path ID: 4-octet identifier that identifies a path (encoded in the ERO/RRO) within the set of multiple paths under the PCEP LSP. Value 0x0 is reserved to indicate the absence of a Path ID. The value of 0x0 MAY be used when this Path is not being referenced by any other path and the allocation of a Path ID is not necessary.

TLVs that may be included in the PATH-ATTRIB object are described in the following sections. Other optional TLVs could be defined by future documents to be included within the PATH-ATTRIB object body.

4.3. Multipath Weight TLV

We define the MULTIPATH-WEIGHT TLV that MAY be present in the PATH-ATTRIB object.

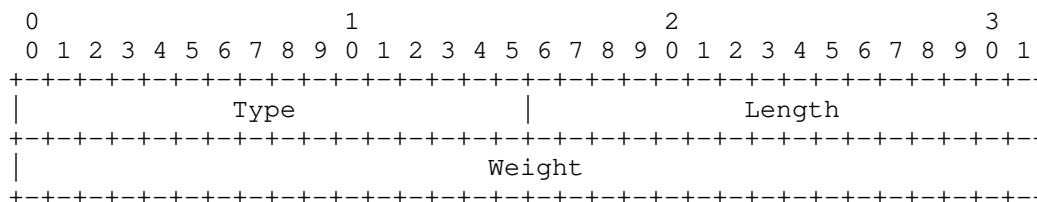


Figure 3: MULTIPATH-WEIGHT TLV format

Type: TBD3 for "MULTIPATH-WEIGHT" TLV.

Length: 4.

Weight: weight of this path within the multipath, if W-ECMP is desired. The fraction of flows a specific ERO/RRO carries is derived from the ratio of its weight to the sum of all other multipath ERO/RRO weights.

When the MULTIPATH-WEIGHT TLV is absent from the PATH-ATTRIB object, or the PATH-ATTRIB object is absent from the <intended-path>/<actual-path>, then the Weight of the corresponding path is taken to be "1".

4.4. Multipath Backup TLV

This document introduces a new MULTIPATH-BACKUP TLV that is optional and can be present in the PATH-ATTRIB object.

This TLV is used to indicate the presence of a backup path that is used for protection in case of failure of the primary path. The format of the MULTIPATH-BACKUP TLV is:

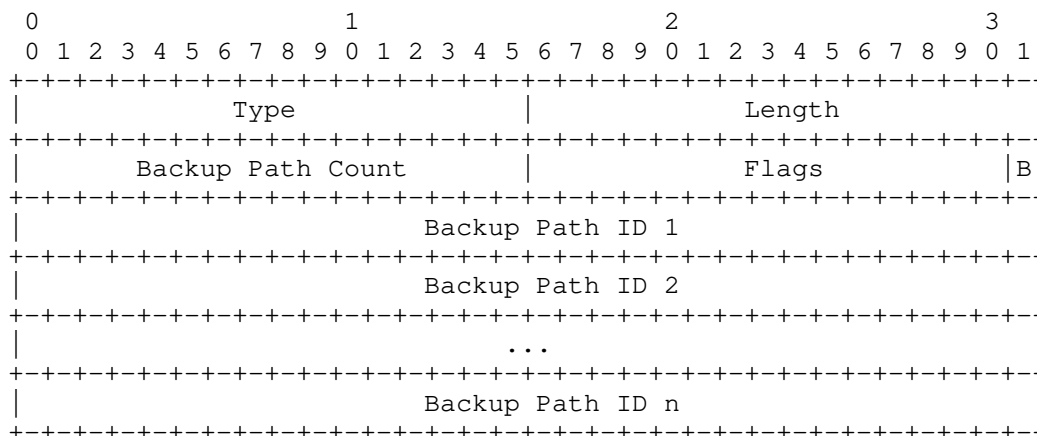


Figure 4: MULTIPATH-BACKUP TLV format

Type: TBD4 for "MULTIPATH-BACKUP" TLV

Length: $4 + (N * 4)$ (where N is the Backup Path Count)

Backup Path Count: Number of backup path(s).

B: If set, indicates a pure backup path. This is a path that only carries rerouted traffic after the protected path fails. If this flag is not set, or if the MULTIPATH-BACKUP TLV is absent, then the path is assumed to be primary that carries normal traffic.

Backup Path ID(s): a series of 4-octet identifier(s) that identify the backup path(s) in the set that protect this primary path.

4.5. Multipath Opposite Direction Path TLV

This document introduces a new MULTIPATH-OPPDIR-PATH TLV that is optional and can be present in the PATH-ATTRIB object.

This TLV is used to indicate whether the given path is a forward path or a reverse path in its PCEP Tunnel, as well as give information about the opposite-direction path(s) of the given path.

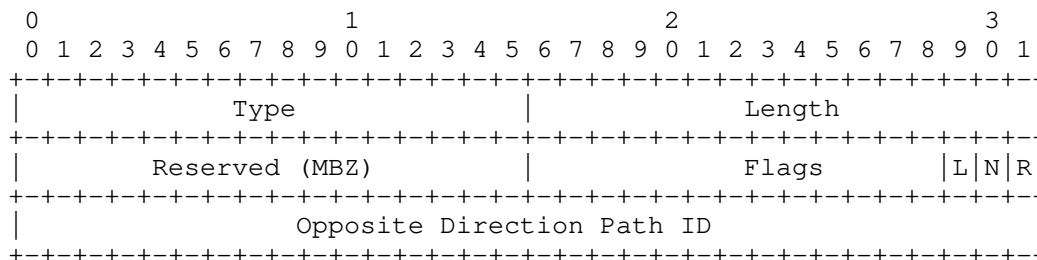


Figure 5: MULTIPATH-OPPDIR-PATH TLV format

Type: TBD9 for "MULTIPATH-OPPDIR-PATH" TLV

Length: 16.

R (Reverse path): If set, indicates this path is reverse, i.e., it originates on the Tunnel destination and terminates on the Tunnel source (usually the PCC headend itself). Paths with this flag set MUST NOT be installed into forwarding, they serve only informational purposes.

N (Node co-routed): If set, indicates this path is node co-routed with its opposite direction path, specified in this TLV. Two opposite direction paths are node co-routed if they traverse the same nodes, but MAY traverse different links.

L (Link co-routed): If set, indicates this path is link co-routed with its opposite directions path, specified in this TLV. Two opposite direction paths are link co-routed if they traverse the same links (but in the opposite directions).

Opposite Direction Path ID: Identifies a path that goes in the opposite direction to this path. If no such path exists, then this field MUST be set to 0x0, which is reserved to indicate the absence of a Path ID.

Multiple instances of this TLV present in the same PATH-ATTRIB object indicate that there are multiple opposite-direction paths corresponding to the given path. This allows for many-to-many relationship among the paths of two opposite direction Tunnels.

Whenever path A references another path B as being the opposite-direction path, then path B typically also reference path A as its own opposite-direction path.

See Section 7.4 for an example of usage.

4.6. Composite Candidate Path

SR Policy Architecture [I-D.ietf-spring-segment-routing-policy] defines the concept of a Composite Candidate Path. Unlike a Non-Composite Candidate Path, which contains Segment Lists, the Composite Candidate Path contains Colors of other policies. The traffic that is steered into a Composite Candidate Path is split among the policies that are identified by the Colors contained in the Composite Candidate Path. The split can be either ECMP or UCMP by adjusting the weight of each color in the Composite Candidate Path, in the same manner as the weight of each Segment List in the Non-Composite Candidate Path is adjusted.

To signal the Composite Candidate Path, we make use of the COLOR TLV, defined in [I-D.draft-rajagopalan-pce-pcep-color]. For a Composite Candidate Path, the COLOR TLV is included in the PATH-ATTRIB Object, thus allowing each Composite Candidate Path to do ECMP/UCMP among SR Policies or Tunnels identified by its constituent Colors. Only one COLOR TLV SHOULD be included into the PATH-ATTRIB object. If multiple COLOR TLVs are contained in the PATH-ATTRIB object, only the first one MUST be processed and the others SHOULD be ignored.

An empty ERO object MUST be included as per the existing RBNF, i.e., ERO MUST contain no sub-objects. If the head-end receives a non-empty ERO, then it MUST send PCErr message with Error-Type 19 ("Invalid Operation") and Error-Value = TBD8 ("Non-empty path").

See Section 7.3 for an example of the encoding.

5. Operation

When the PCC wants to indicate to the PCE that it wants to get multipaths for a PCEP Tunnel, instead of a single path, it can do (1) or both (1) and (2) of the following:

(1) Send the MULTIPATH-CAP TLV in the OPEN object during session establishment. This applies to all PCEP Tunnels on the PCC, unless overridden by PCEP Tunnel specific information.

(2) Additionally send the MULTIPATH-CAP TLV in the LSP object for a particular PCEP Tunnel in the PCrpt or PCReq message. This applies to the specified PCEP Tunnel and overrides the information from the OPEN object.

When PCE computes the path for a PCEP Tunnel, it MUST NOT return more multipaths than the corresponding value of "Number of Multipaths" from the MULTIPATH-CAP TLV. If this TLV is absent (from both OPEN and LSP objects), then the "Number of Multipaths" is assumed to be 1.

If the PCE supports this standard, then it MUST include the MULTIPATH-CAP TLV in the OPEN object. This tells the PCC that it can report multiple ERO/RRO objects per PCEP Tunnel to this PCE. If the PCE does not include the MULTIPATH-CAP TLV in the OPEN object, then the PCC MUST assume that the PCE does not support this standard and fall back to reporting only a single ERO/RRO. The PCE MUST NOT include MULTIPATH-CAP TLV in the LSP object in any other PCEP message towards the PCC and the PCC MUST ignore it if received.

The Path ID of each ERO/RRO MUST be unique within that LSP. If a PCEP speaker detects that there are two paths with the same Path ID, then the PCEP speaker SHOULD send PCErr message with Error-Type = 1 ("Reception of an invalid object") and Error-Value = TBD5 ("Conflicting Path ID").

5.1. Signaling Multiple Paths for Loadbalancing

The PATH-ATTRIB object can be used to signal multiple path(s) and indicate (un)equal loadbalancing amongst the set of multipaths. In this case, the PATH-ATTRIB is populated for each ERO as follows:

1. The PCE assigns a unique Path ID to each ERO path and populates it inside the PATH-ATTRIB object. The Path ID is unique within the context of a PLSP or PCEP Tunnel.
2. The MULTIPATH-WEIGHT TLV MAY be carried inside the PATH-ATTRIB object. A weight is populated to reflect the relative loadshare that is to be carried by the path. If the MULTIPATH-WEIGHT is not carried inside a PATH-ATTRIB object, the default weight 1 MUST be assumed when computing the loadshare.
3. The fraction of flows carried by a specific primary path is derived from the ratio of its weight to the sum of all other multipath weights.

5.2. Signaling Multiple Paths for Protection

The PATH-ATTRIB object can be used to describe a set of backup path(s) protecting a primary path within a PCEP Tunnel. In this case, the PATH-ATTRIB is populated for each ERO as follows:

1. The PCE assigns a unique Path ID to each ERO path and populates it inside the PATH-ATTRIB object. The Path ID is unique within the context of a PLSP or PCEP Tunnel.
2. The MULTIPATH-BACKUP TLV MUST be added inside the PATH-ATTRIB object for each ERO that is protected. The backup path ID(s) are populated in the MULTIPATH-BACKUP TLV to reflect the set of backup path(s) protecting the primary path. The Length field and Backup Path Number in the MULTIPATH-BACKUP are updated according to the number of backup path ID(s) included.
3. The MULTIPATH-BACKUP TLV MAY be added inside the PATH-ATTRIB object for each ERO that is unprotected. In this case, MULTIPATH-BACKUP does not carry any backup path IDs in the TLV. If the path acts as a pure backup - i.e. the path only carries rerouted traffic after the protected path(s) fail- then the B flag MUST be set.

Note that if a given path has the B-flag set, then there MUST be some other path within the same LSP that uses the given path as a backup. If this condition is violated, then the PCEP speaker SHOULD send a PCError message with Error-Type = 10 ("Reception of an invalid object") and Error-Value = TBD6 ("No primary path for pure backup").

Note that a given PCC may not support certain backup combinations, such as a backup path that is itself protected by another backup path, etc. If a PCC is not able to implement a requested backup scenario, the PCC SHOULD send a PCError message with Error-Type = 19 ("Invalid Operation") and Error-Value = TBD7 ("Not supported path backup").

6. PCEP Message Extensions

The RBNF of PCReq, PCRep, PCRpt, PCUpd and PCInit messages currently use a combination of <intended-path> and/or <actual-path>. As specified in Section 6.1 of [RFC8231], <intended-path> is represented by the ERO object and <actual-path> is represented by the RRO object:

```
<intended-path> ::= <ERO>
```

```
<actual-path> ::= <RRO>
```

In this standard, we extend these two elements to allow multiple ERO/RRO objects to be present in the <intended-path>/<actual-path>:

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

```
<actual-path> ::= (<RRO> |
                  (<PATH-ATTRIB><RRO>)
                  [<actual-path>])
```

7. Examples

7.1. SR Policy Candidate-Path with Multiple Segment-Lists

Consider the following sample SR Policy, taken from [I-D.ietf-spring-segment-routing-policy].

```
SR policy POL1 <headend, color, endpoint>
  Candidate-path CP1 <protocol-origin = 20, originator =
100:1.1.1.1, discriminator = 1>
    Preference 200
    Weight W1, SID-List1 <SID11...SID1i>
    Weight W2, SID-List2 <SID21...SID2j>
  Candidate-path CP2 <protocol-origin = 20, originator =
100:2.2.2.2, discriminator = 2>
    Preference 100
    Weight W3, SID-List3 <SID31...SID3i>
    Weight W4, SID-List4 <SID41...SID4j>
```

As specified in [I-D.ietf-pce-segment-routing-policy-cp], CP1 and CP2 are signaled as separate state-report elements and each has a unique PLSP-ID, assigned by the PCC. Let us assign PLSP-ID 100 to CP1 and PLSP-ID 200 to CP2.

The state-report for CP1 can be encoded as:

```
<state-report> =  
  <LSP PLSP_ID=100>  
  <ASSOCIATION>  
  <END-POINT>  
  <PATH-ATTRIB Path_ID=1 <WEIGHT-TLV Weight=W1>>  
  <ERO SID-List1>  
  <PATH-ATTRIB Path_ID=2 <WEIGHT-TLV Weight=W2>>  
  <ERO SID-List2>
```

The state-report for CP2 can be encoded as:

```
<state-report> =  
  <LSP PLSP_ID=200>  
  <ASSOCIATION>  
  <END-POINT>  
  <PATH-ATTRIB Path_ID=1 <WEIGHT-TLV Weight=W3>>  
  <ERO SID-List3>  
  <PATH-ATTRIB Path_ID=2 <WEIGHT-TLV Weight=W4>>  
  <ERO SID-List4>
```

The above sample state-report elements only specify the minimum mandatory objects, of course other objects like SRP, LSPA, METRIC, etc., are allowed to be inserted.

Note that the syntax

```
<PATH-ATTRIB Path_ID=1 <WEIGHT-TLV Weight=W1>>
```

, simply means that this is PATH-ATTRIB object with Path ID field set to "1" and with a MULTIPATH-WEIGHT TLV carrying weight of "W1".

7.2. Two Primary Paths Protected by One Backup Path

Suppose there are 3 paths: A, B, C. Where A,B are primary and C is to be used only when A or B fail. Suppose the Path IDs for A, B, C are respectively 1, 2, 3. This would be encoded in a state-report as:

```

<state-report> =
  <LSP>
  <ASSOCIATION>
  <END-POINT>
  <PATH-ATTRIB Path_ID=1 <BACKUP-TLV B=0, Backup_Paths=[3]>>
  <ERO A>
  <PATH-ATTRIB Path_ID=2 <BACKUP-TLV B=0, Backup_Paths=[3]>>
  <ERO B>
  <PATH-ATTRIB Path_ID=3 <BACKUP-TLV B=1, Backup_Paths=[]>>
  <ERO C>

```

Note that the syntax

```

<PATH-ATTRIB Path_ID=1 <BACKUP-TLV B=0, Backup_Paths=[3]>>

```

, simply means that this is PATH-ATTRIB object with Path ID field set to "1" and with a MULTIPATH-BACKUP TLV that has B-flag cleared and contains a single backup path with Backup Path ID of 3.

7.3. Composite Candidate Path

Consider the following Composite Candidate Path, taken from [I-D.ietf-spring-segment-routing-policy].

```

SR policy POL100 <headend = H1, color = 100, endpoint = E1>
  Candidate-path CP1 <protocol-origin = 20, originator =
100:1.1.1.1, discriminator = 1>
    Preference 200
    Weight W1, SR policy <color = 1>
    Weight W2, SR policy <color = 2>

```

This is signaled in PCEP as:

```

<LSP PLSP_ID=100>
  <ASSOCIATION>
  <END-POINT>
  <PATH-ATTRIB Path_ID=1
    <WEIGHT-TLV Weight=W1>
    <COLOR-TLV Color=1>>
  <ERO (empty)>
  <PATH-ATTRIB Path_ID=2
    <WEIGHT-TLV Weight=W2>
    <COLOR-TLV Color=2>>
  <ERO (empty)>

```

7.4. Opposite Direction Tunnels

Consider the two opposite-direction SR Policies between end-points H1 and E1.

```
SR policy POL1 <headend = H1, color, endpoint = E1>
  Candidate-path CP1
    Preference 200
    Bidirectional Association = A1
    SID-List = <H1,M1,M2,E1>
    SID-List = <H1,M3,M4,E1>
  Candidate-path CP2
    Preference 100
    Bidirectional Association = A2
    SID-List = <H1,M5,M6,E1>
    SID-List = <H1,M7,M8,E1>

SR policy POL2 <headend = E1, color, endpoint = H1>
  Candidate-path CP1
    Preference 200
    Bidirectional Association = A1
    SID-List = <E1,M2,M1,H1>
    SID-List = <E1,M4,M3,H1>
  Candidate-path CP2
    Preference 100
    Bidirectional Association = A2
    SID-List = <E1,M6,M5,H1>
```

The state-report for POL1, CP1 can be encoded as:

```
<state-report> =
  <LSP PLSP_ID=100>
  <BIDIRECTIONAL ASSOCIATION = A1>
  <PATH-ATTRIB PathID=1>
    <OPPDIR-PATH-TLV R-flag=0 OppositePathID=3>>
  <ERO <H1,M1,M2,E1>>
  <PATH-ATTRIB PathID=2>
    <OPPDIR-PATH-TLV R-flag=0 OppositePathID=4>>
  <ERO <H1,M3,M4,E1>>
  <PATH-ATTRIB PathID=3>
    <OPPDIR-PATH-TLV R-flag=1 OppositePathID=1>>
  <ERO <E1,M2,M1,H1>>
  <PATH-ATTRIB PathID=4>
    <OPPDIR-PATH-TLV R-flag=1 OppositePathID=2>>
  <ERO <E1,M4,M3,H1>>
```

The state-report for POL1, CP2 can be encoded as:

```

<state-report> =
  <LSP PLSP_ID=200>
  <BIDIRECTIONAL ASSOCIATION = A2>
  <PATH-ATTRIB PathID=1
    <OPPDIR-PATH-TLV R-flag=0 OppositePathID=3>>
  <ERO <H1,M5,N6,E1>>
  <PATH-ATTRIB PathID=2
    <OPPDIR-PATH-TLV R-flag=0 OppositePathID=0>>
  <ERO <H1,M7,M8,E1>>
  <PATH-ATTRIB PathID=3
    <OPPDIR-PATH-TLV R-flag=1 OppositePathID=1>>
  <ERO <E1,M6,M5,H1>>

```

The state-report for POL2, CP1 can be encoded as:

```

<state-report> =
  <LSP PLSP_ID=100>
  <BIDIRECTIONAL ASSOCIATION = A1>
  <PATH-ATTRIB PathID=1
    <OPPDIR-PATH-TLV R-flag=0 OppositePathID=3>>
  <ERO <E1,M2,M1,H1>>
  <PATH-ATTRIB PathID=2
    <OPPDIR-PATH-TLV R-flag=0 OppositePathID=4>>
  <ERO <E1,M4,M3,H1>>
  <PATH-ATTRIB PathID=3
    <OPPDIR-PATH-TLV R-flag=1 OppositePathID=1>>
  <ERO <H1,M1,M2,E1>>
  <PATH-ATTRIB PathID=4
    <OPPDIR-PATH-TLV R-flag=1 OppositePathID=2>>
  <ERO <H1,M3,M4,E1>>

```

The state-report for POL2, CP2 can be encoded as:

```

<state-report> =
  <LSP PLSP_ID=200>
  <BIDIRECTIONAL ASSOCIATION = A2>
  <PATH-ATTRIB PathID=1
    <OPPDIR-PATH-TLV R-flag=0 OppositePathID=3>>
  <ERO <E1,M6,M5,H1>>
  <PATH-ATTRIB PathID=2
    <OPPDIR-PATH-TLV R-flag=1 OppositePathID=0>>
  <ERO <H1,M7,M8,E1>>
  <PATH-ATTRIB PathID=3
    <OPPDIR-PATH-TLV R-flag=1 OppositePathID=1>>
  <ERO <H1,M5,N6,E1>>

```

8. IANA Considerations

8.1. PCEP Object

IANA is requested to make the assignment of a new value for the existing "PCEP Objects" registry as follows:

Object-Class Value	Name	Object-Type Value	Reference
TBD2	PATH-ATTRIB	1	This document

8.2. PCEP TLV

IANA is requested to make the assignment of a new value for the existing "PCEP TLV Type Indicators" registry as follows:

TLV Type Value	TLV Name	Reference
TBD1	MULTIPATH-CAP	This document
TBD3	MULTIPATH-WEIGHT	This document
TBD4	MULTIPATH-BACKUP	This document
TBD9	MULTIPATH-OPPDIR-PATH	This document

8.3. PCEP-Error Object

IANA is requested to make the assignment of a new value for the existing "PCEP-ERROR Object Error Types and Values" sub-registry of the PCEP Numbers registry for the following errors:

Error-Type	Error-Value	Reference
10	TBD5 - Conflicting Path ID	This document
10	TBD6 - No primary path for pure backup	This document
19	TBD7 - Not supported path backup	This document
19	TBD8 - Non-empty path	This document

8.4. Flags in the Multipath Capability TLV

IANA is requested to create a new sub-registry to manage the Flag field of the MULTIPATH-CAP TLV, called "Flags in MULTIPATH-CAP TLV".

Bit	Description	Reference
0-12	Unassigned	This document
13	O-flag: support for processing MULTIPATH-OPPDIR-PATH TLV	This document
14	B-flag: support for processing MULTIPATH-BACKUP TLV	This document
15	W-flag: support for processing MULTIPATH-WEIGHT TLV	This document

8.5. Flags in the Path Attribute Object

IANA is requested to create a new sub-registry to manage the Flag field of the PATH-ATTRIBUTE object, called "Flags in PATH-ATTRIBUTE Object".

Bit	Description	Reference
0-12	Unassigned	This document
13-15	O-flag: Operational state	This document

8.6. Flags in the Multipath Backup TLV

IANA is requested to create a new sub-registry to manage the Flag field of the MULTIPATH-BACKUP TLV, called "Flags in MULTIPATH-BACKUP TLV".

Bit	Description	Reference
0-14	Unassigned	This document
15	B-flag: Pure backup	This document

8.7. Flags in the Multipath Opposite Direction Path TLV

IANA is requested to create a new sub-registry to manage the flag fields of the MULTIPATH-OPPDIR-PATH TLV, called "Flags in the MULTIPATH-OPPDIR-PATH TLV".

Bit	Description	Reference
0-12	Unassigned	This document
13	L-flag: Link co-routed	This document
14	N-flag: Node co-routed	This document
15	R-flag: Reverse path	This document

9. Security Considerations

None at this time.

10. Acknowledgement

Thanks to Dhruv Dhody for ideas and discussion.

11. Contributors

Andrew Stone
Nokia

Email: andrew.stone@nokia.com

Gyan Mishra
Verizon Inc.

Email: gyan.s.mishra@verizon.com

12. References

12.1. Normative References

- [I-D.draft-rajagopalan-pce-pcep-color]
Rajagopalan, B., Beeram, V. P., Peng, S., Xiong, Q., Koldychev, M., and G. Mishra, "Path Computation Element Protocol (PCEP) Extension for Color", Work in Progress, Internet-Draft, draft-rajagopalan-pce-pcep-color-00, 25 October 2021, <<https://www.ietf.org/archive/id/draft-rajagopalan-pce-pcep-color-00.txt>>.
- [I-D.ietf-pce-segment-routing-policy-cp]
Koldychev, M., Sivabalan, S., Barth, C., Peng, S., and H. Bidgoli, "PCEP extension to support Segment Routing Policy Candidate Paths", Work in Progress, Internet-Draft, draft-ietf-pce-segment-routing-policy-cp-06, 22 October 2021, <<https://www.ietf.org/archive/id/draft-ietf-pce-segment-routing-policy-cp-06.txt>>.
- [I-D.ietf-spring-segment-routing-policy]
Filsfils, C., Talaulikar, K., Voyer, D., Bogdanov, A., and P. Mattes, "Segment Routing Policy Architecture", Work in Progress, Internet-Draft, draft-ietf-spring-segment-routing-policy-14, 25 October 2021, <<https://www.ietf.org/archive/id/draft-ietf-spring-segment-routing-policy-14.txt>>.
- [I-D.koldychev-pce-operational]
Koldychev, M., Sivabalan, S., Peng, S., Achaval, D., and H. Kotni, "PCEP Operational Clarification", Work in Progress, Internet-Draft, draft-koldychev-pce-operational-04, 19 August 2021, <<https://www.ietf.org/archive/id/draft-koldychev-pce-operational-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>>.
- [RFC8174] Leiba, B., "Ambiguity of Uppercase vs Lowercase in RFC 2119 Key Words", BCP 14, RFC 8174, DOI 10.17487/RFC8174, May 2017, <<https://www.rfc-editor.org/info/rfc8174>>.

- [RFC8231] Crabbe, E., Minei, I., Medved, J., and R. Varga, "Path Computation Element Communication Protocol (PCEP) Extensions for Stateful PCE", RFC 8231, DOI 10.17487/RFC8231, September 2017, <<https://www.rfc-editor.org/info/rfc8231>>.
- [RFC8281] Crabbe, E., Minei, I., Sivabalan, S., and R. Varga, "Path Computation Element Communication Protocol (PCEP) Extensions for PCE-Initiated LSP Setup in a Stateful PCE Model", RFC 8281, DOI 10.17487/RFC8281, December 2017, <<https://www.rfc-editor.org/info/rfc8281>>.
- [RFC8664] Sivabalan, S., Filsfils, C., Tantsura, J., Henderickx, W., and J. Hardwick, "Path Computation Element Communication Protocol (PCEP) Extensions for Segment Routing", RFC 8664, DOI 10.17487/RFC8664, December 2019, <<https://www.rfc-editor.org/info/rfc8664>>.

12.2. Informative References

- [RFC4655] Farrel, A., Vasseur, J.-P., and J. Ash, "A Path Computation Element (PCE)-Based Architecture", RFC 4655, DOI 10.17487/RFC4655, August 2006, <<https://www.rfc-editor.org/info/rfc4655>>.
- [RFC8745] Ananthakrishnan, H., Sivabalan, S., Barth, C., Minei, I., and M. Negi, "Path Computation Element Communication Protocol (PCEP) Extensions for Associating Working and Protection Label Switched Paths (LSPs) with Stateful PCE", RFC 8745, DOI 10.17487/RFC8745, March 2020, <<https://www.rfc-editor.org/info/rfc8745>>.

Authors' Addresses

Mike Koldychev
Cisco Systems, Inc.

Email: mkoldych@cisco.com

Siva Sivabalan
Ciena Corporation

Email: ssivabal@ciena.com

Tarek Saad
Juniper Networks, Inc.

Email: tsaad@juniper.net

Vishnu Pavan Beeram
Juniper Networks, Inc.

Email: vbeeram@juniper.net

Hooman Bidgoli
Nokia

Email: hooman.bidgoli@nokia.com

Bhupendra Yadav
Ciena

Email: byadav@ciena.com

Shuping Peng
Huawei Technologies

Email: pengshuping@huawei.com

PCE Working Group
Internet-Draft
Intended status: Standards Track
Expires: 1 October 2022

M. Koldychev
Cisco Systems, Inc.
S. Sivabalan
Ciena Corporation
T. Saad
V. Beeram
Juniper Networks, Inc.
H. Bidgoli
Nokia
B. Yadav
Ciena
S. Peng
Huawei Technologies
G. Mishra
Verizon Inc.
30 March 2022

PCEP Extensions for Signaling Multipath Information
draft-ietf-pce-multipath-05

Abstract

Path computation algorithms are not limited to return a single optimal path. Multiple paths may exist that satisfy the given objectives and constraints. This document defines a mechanism to encode multiple paths for a single set of objectives and constraints. This is a generic PCEP mechanism, not specific to any path setup type or dataplane. The mechanism is applicable to both stateless and stateful PCEP.

Status of This Memo

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

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at <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 1 October 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. Terminology	4
2.1. Terms and Abbreviations	4
3. Motivation	4
3.1. Signaling Multiple Segment-Lists of an SR Candidate-Path	4
3.2. Splitting of Requested Bandwidth	4
3.3. Providing Backup path for Protection	4
3.4. Reverse Path Information	5
4. Protocol Extensions	5
4.1. Multipath Capability TLV	5
4.2. Path Attributes Object	6
4.3. Multipath Weight TLV	7
4.4. Multipath Backup TLV	7
4.5. Multipath Opposite Direction Path TLV	8
4.6. Composite Candidate Path	10
5. Operation	10
5.1. Capability Negotiation	10
5.2. Path ID	11
5.3. Signaling Multiple Paths for Loadbalancing	11
5.4. Signaling Multiple Paths for Protection	12
6. PCEP Message Extensions	13
7. Examples	13
7.1. SR Policy Candidate-Path with Multiple Segment-Lists	13
7.2. Two Primary Paths Protected by One Backup Path	14
7.3. Composite Candidate Path	15
7.4. Opposite Direction Tunnels	15
8. Implementation Status	18
8.1. Cisco Systems	18
8.2. Ciena Corp	18
9. IANA Considerations	18
9.1. PCEP Object	18

9.2. PCEP TLV	19
9.3. PCEP-Error Object	19
9.4. Flags in the Multipath Capability TLV	19
9.5. Flags in the Path Attribute Object	20
9.6. Flags in the Multipath Backup TLV	20
9.7. Flags in the Multipath Opposite Direction Path TLV	21
10. Security Considerations	21
11. Acknowledgement	21
12. Contributors	21
13. References	21
13.1. Normative References	21
13.2. Informative References	23
Authors' Addresses	23

1. Introduction

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

PCEP Extensions for the Stateful PCE Model [RFC8231] describes a set of extensions to PCEP that enable active control of Multiprotocol Label Switching Traffic Engineering (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.

PCEP Extensions for Segment Routing [RFC8664] specifies extensions to the Path Computation Element Protocol (PCEP) that allow a stateful PCE to compute and initiate Traffic Engineering (TE) paths, as well as for a PCC to request a path subject to certain constraint(s) and optimization criteria in SR networks.

Segment Routing Policy for Traffic Engineering [I-D.ietf-spring-segment-routing-policy] details the concepts of SR Policy and approaches to steering traffic into an SR Policy. In particular, it describes the SR candidate-path as a collection of one or more Segment-Lists. The current PCEP standards only allow for signaling of one Segment-List per Candidate-Path. PCEP extension to support Segment Routing Policy Candidate Paths [I-D.ietf-pce-segment-routing-policy-cp] specifically avoids defining how to signal multipath information, and states that this will be defined in another document.

This document defines the required extensions that allow the signaling of multipath information via PCEP.

2. Terminology

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.

2.1. Terms and Abbreviations

The following terms are used in this document:

PCEP Tunnel:

The object identified by the PLSP-ID, see [I-D.koldychev-pce-operational] for more details.

3. Motivation

This extension is motivated by the use-cases described below.

3.1. Signaling Multiple Segment-Lists of an SR Candidate-Path

The Candidate-Path of an SR Policy is the unit of report/update in PCEP, see [I-D.ietf-pce-segment-routing-policy-cp]. Each Candidate-Path can contain multiple Segment-Lists and each Segment-List is encoded by one ERO. However, each PCEP LSP can contain only a single ERO, which prevents us from encoding multiple Segment-Lists within the same SR Candidate-Path.

With the help of the protocol extensions defined in this document, this limitation is overcome.

3.2. Splitting of Requested Bandwidth

A PCC may request a path with 80 Gbps of bandwidth, but all links in the network have only 50 Gbps capacity. The PCE can return two paths, that can together carry 80 Gbps. The PCC can then equally or unequally split the incoming 80 Gbps of traffic among the two paths. Section 4.3 introduces a new TLV that carries the path weight that allows for distribution of incoming traffic on to the multiple paths.

3.3. Providing Backup path for Protection

It is desirable for the PCE to compute and signal to the PCC a backup path that is used to protect a primary path within the multipaths in a given LSP.

Note that [RFC8745] specify the Path Protection association among LSPs. The use of [RFC8745] with multipath is out of scope of this document and is for future study.

When multipath is used, a backup path may protect one or more primary paths. For this reason, primary and backup path identifiers are needed to indicate which backup path(s) protect which primary path(s). Section 4.4 introduces a new TLV that carries the required information.

3.4. Reverse Path Information

Certain applications, such as Circuit Style SR Policy [I-D.schmutzer-pce-cs-sr-policy], require the head-end to know both forward and reverse paths for each of the segment lists of an SR Policy in order to run OAM/PM/BFD protocols on each Segment List as a separate circuit.

4. Protocol Extensions

4.1. Multipath Capability TLV

We define the MULTIPATH-CAP TLV that MAY be present in the OPEN object and/or the LSP object. The purpose of this TLV is two-fold:

1. From PCC: it tells how many multipaths per PCEP Tunnel, the PCC can install in forwarding.
2. From PCE: it tells that the PCE supports this standard and how many multipaths per PCEP Tunnel, the PCE can compute.

Only the first instance of this TLV can be processed, subsequent instances SHOULD be ignored.

Section 5 specify the usage of this TLV with Open message (within the OPEN object) and other PCEP messages (within the LSP object).

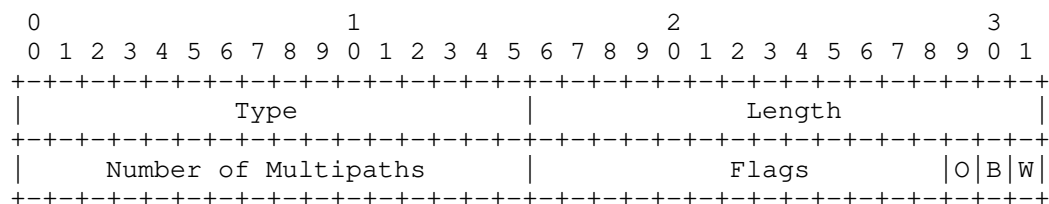


Figure 1: MULTIPATH-CAP TLV format

Type: TBD1 for "MULTIPATH-CAP" TLV.

TLVs that may be included in the PATH-ATTRIB object are described in the following sections. Other optional TLVs could be defined by future documents to be included within the PATH-ATTRIB object body.

4.3. Multipath Weight TLV

We define the MULTIPATH-WEIGHT TLV that MAY be present in the PATH-ATTRIB object.

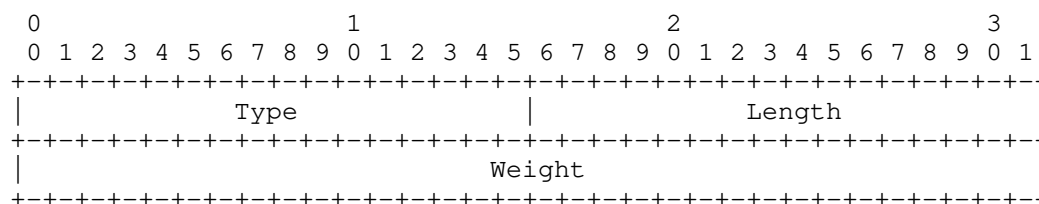


Figure 3: MULTIPATH-WEIGHT TLV format

Type: TBD3 for "MULTIPATH-WEIGHT" TLV.

Length: 4.

Weight: weight of this path within the multipath, if W-ECMP is desired. The fraction of flows a specific ERO/RRO carries is derived from the ratio of its weight to the sum of all other multipath ERO/RRO weights.

When the MULTIPATH-WEIGHT TLV is absent from the PATH-ATTRIB object, or the PATH-ATTRIB object is absent from the <intended-path>/<actual-path>, then the Weight of the corresponding path is taken to be "1".

4.4. Multipath Backup TLV

This document introduces a new MULTIPATH-BACKUP TLV that MAY be present in the PATH-ATTRIB object.

This TLV is used to indicate the presence of a backup path that is used for protection in case of failure of the primary path. The format of the MULTIPATH-BACKUP TLV is:

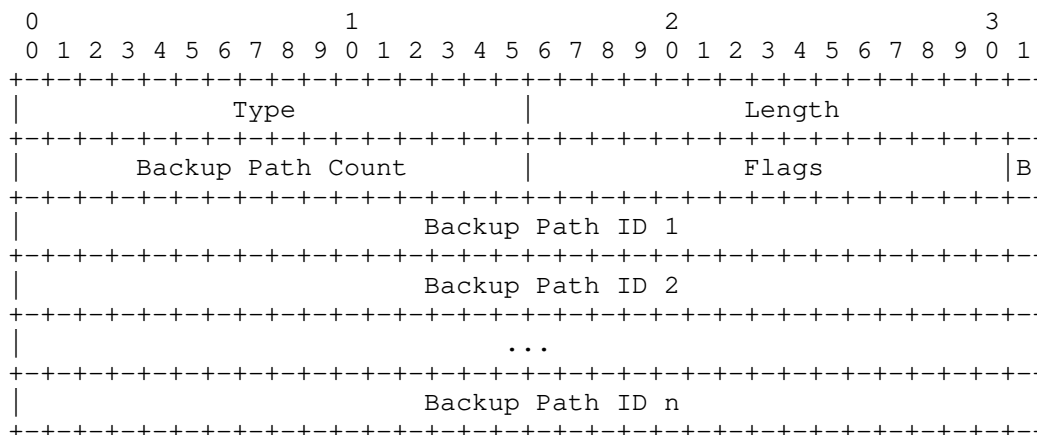


Figure 4: MULTIPATH-BACKUP TLV format

Type: TBD4 for "MULTIPATH-BACKUP" TLV

Length: $4 + (N * 4)$ (where N is the Backup Path Count)

Backup Path Count: Number of backup path(s).

B: If set, indicates a pure backup path. This is a path that only carries rerouted traffic after the protected path fails. If this flag is not set, or if the MULTIPATH-BACKUP TLV is absent, then the path is assumed to be primary that carries normal traffic.

Backup Path ID(s): a series of 4-octet identifier(s) that identify the backup path(s) in the set that protect this primary path.

4.5. Multipath Opposite Direction Path TLV

This document introduces a new MULTIPATH-OPPDIR-PATH TLV that MAY be present in the PATH-ATTRIB object. This TLV encodes a many-to-many mapping between forward and reverse paths within a PCEP Tunnel.

Many-to-many mapping means that a single forward path MAY map to multiple reverse paths and conversely that a single reverse path MAY map to multiple forward paths. Many-to-many mapping can happen for an SR Policy, when a Segment List contains Node Segment(s) which traverse parallel links at the midpoint. The reverse of this Segment List may not be able to be expressed as a single Reverse Segment List, but need to return multiple Reverse Segment Lists to cover all the parallel links at the midpoint.

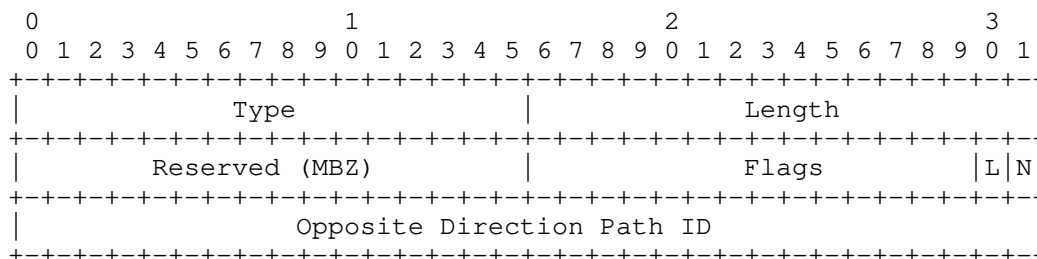


Figure 5: MULTIPATH-OPPDIR-PATH TLV format

Type: TBD9 for "MULTIPATH-OPPDIR-PATH" TLV

Length: 16.

N (Node co-routed): If set, indicates this path is node co-routed with its opposite direction path, specified in this TLV. Two opposite direction paths are node co-routed if they traverse the same nodes, but MAY traverse different links.

L (Link co-routed): If set, indicates this path is link co-routed with its opposite directions path, specified in this TLV. Two opposite direction paths are link co-routed if they traverse the same links (but in the opposite directions).

Opposite Direction Path ID: Identifies a path that goes in the opposite direction to this path. If no such path exists, then this field MUST be set to 0x0, which is reserved to indicate the absence of a Path ID.

Multiple instances of this TLV present in the same PATH-ATTRIB object indicate that there are multiple opposite-direction paths corresponding to the given path. This allows for many-to-many relationship among the paths of two opposite direction Tunnels.

Whenever path A references another path B as being the opposite-direction path, then path B typically also reference path A as its own opposite-direction path.

See Section 7.4 for an example of usage.

4.6. Composite Candidate Path

SR Policy Architecture [I-D.ietf-spring-segment-routing-policy] defines the concept of a Composite Candidate Path. Unlike a Non-Composite Candidate Path, which contains Segment Lists, the Composite Candidate Path contains Colors of other policies. The traffic that is steered into a Composite Candidate Path is split among the policies that are identified by the Colors contained in the Composite Candidate Path. The split can be either ECMP or UCMP by adjusting the weight of each color in the Composite Candidate Path, in the same manner as the weight of each Segment List in the Non-Composite Candidate Path is adjusted.

To signal the Composite Candidate Path, we make use of the COLOR TLV, defined in [I-D.draft-rajagopalan-pce-pcep-color]. For a Composite Candidate Path, the COLOR TLV is included in the PATH-ATTRIB Object, thus allowing each Composite Candidate Path to do ECMP/UCMP among SR Policies or Tunnels identified by its constituent Colors. Only one COLOR TLV SHOULD be included into the PATH-ATTRIB object. If multiple COLOR TLVs are contained in the PATH-ATTRIB object, only the first one MUST be processed and the others SHOULD be ignored.

An empty ERO object MUST be included as per the existing RBNF, i.e., ERO MUST contain no sub-objects. If the head-end receives a non-empty ERO, then it MUST send PCError message with Error-Type 19 ("Invalid Operation") and Error-Value = TBD8 ("Non-empty path").

See Section 7.3 for an example of the encoding.

5. Operation

5.1. Capability Negotiation

When the PCC wants to indicate to the PCE that it wants to get multipaths for a PCEP Tunnel, instead of a single path, it can do either (1) or both (1) and (2) of the following:

(1) Send the MULTIPATH-CAP TLV in the OPEN object during session establishment. This applies to all PCEP Tunnels on the PCC, unless overridden by PCEP Tunnel specific information.

(2) Additionally send the MULTIPATH-CAP TLV in the LSP object for a particular PCEP Tunnel in the PCRpt or PCReq message. This applies to the specified PCEP Tunnel and overrides the information from the OPEN object.

When PCE computes the path for a PCEP Tunnel, it MUST NOT return more multipaths than the corresponding value of "Number of Multipaths" from the MULTIPATH-CAP TLV. If this TLV is absent (from both OPEN and LSP objects), then the "Number of Multipaths" is assumed to be 1.

If the PCE supports this standard, then it MUST include the MULTIPATH-CAP TLV in the OPEN object. This tells the PCC that it can report multiple ERO/RRO objects per PCEP Tunnel to this PCE. If the PCE does not include the MULTIPATH-CAP TLV in the OPEN object, then the PCC MUST assume that the PCE does not support this standard and fall back to reporting only a single ERO/RRO.

5.2. Path ID

The Path ID uniquely identifies a Path within the context of a PCEP Tunnel. Note that when the PCEP Tunnel is an SR Policy Candidate Path, the Paths within that tunnel are the Segment Lists of that Candidate Path.

Value 0x0 is reserved to indicate the absence of a Path ID. The value of 0x0 MAY be used when this Path is not being referenced and the allocation of a Path ID is not necessary.

Path IDs are allocated by the PCEP peer that currently owns the Tunnel. If the Tunnel is delegated to the PCE, then the PCE allocates the Path IDs and sends them in the PCReply/PCUpd/PCInit messages. If the Tunnel is locally computed on the PCC, then the PCC allocates the Path IDs and sends them in the PCReq/PCRpt messages.

If a PCEP speaker detects that there are two Paths with the same Path ID, then the PCEP speaker SHOULD send PCErr message with Error-Type = 1 ("Reception of an invalid object") and Error-Value = TBD5 ("Conflicting Path ID").

5.3. Signaling Multiple Paths for Loadbalancing

The PATH-ATTRIB object can be used to signal multiple path(s) and indicate (un)equal loadbalancing amongst the set of multipaths. In this case, the PATH-ATTRIB is populated for each ERO as follows:

1. The PCE assigns a unique Path ID to each ERO path and populates it inside the PATH-ATTRIB object. The Path ID is unique within the context of a PLSP or PCEP Tunnel.

2. The MULTIPATH-WEIGHT TLV MAY be carried inside the PATH-ATTRIB object. A weight is populated to reflect the relative loadshare that is to be carried by the path. If the MULTIPATH-WEIGHT is not carried inside a PATH-ATTRIB object, the default weight 1 MUST be assumed when computing the loadshare.
3. The fraction of flows carried by a specific primary path is derived from the ratio of its weight to the sum of all other multipath weights.

5.4. Signaling Multiple Paths for Protection

The PATH-ATTRIB object can be used to describe a set of backup path(s) protecting a primary path within a PCEP Tunnel. In this case, the PATH-ATTRIB is populated for each ERO as follows:

1. The PCE assigns a unique Path ID to each ERO path and populates it inside the PATH-ATTRIB object. The Path ID is unique within the context of a PLSP or PCEP Tunnel.
2. The MULTIPATH-BACKUP TLV MAY be added inside the PATH-ATTRIB object for each ERO that is protected. The backup path ID(s) are populated in the MULTIPATH-BACKUP TLV to reflect the set of backup path(s) protecting the primary path. The Length field and Backup Path Number in the MULTIPATH-BACKUP are updated according to the number of backup path ID(s) included.
3. The MULTIPATH-BACKUP TLV MAY be added inside the PATH-ATTRIB object for each ERO that is unprotected. In this case, MULTIPATH-BACKUP does not carry any backup path IDs in the TLV. If the path acts as a pure backup - i.e. the path only carries rerouted traffic after the protected path(s) fail- then the B flag MUST be set.

Note that primary paths which do not include the MULTIPATH-BACKUP TLV are assumed to be protected by all the backup paths. I.e., omitting the TLV is equivalent to including the TLV with all the backup path IDs filled in.

Note that a given PCC may not support certain backup combinations, such as a backup path that is itself protected by another backup path, etc. If a PCC is not able to implement a requested backup scenario, the PCC SHOULD send a PCErr message with Error-Type = 19 ("Invalid Operation") and Error-Value = TBD7 ("Not supported path backup").

6. PCEP Message Extensions

The RBNF of PCReq, PCRep, PCRpt, PCUpd and PCInit messages currently use a combination of <intended-path> and/or <actual-path>. As specified in Section 6.1 of [RFC8231], <intended-path> is represented by the ERO object and <actual-path> is represented by the RRO object:

```
<intended-path> ::= <ERO>
```

```
<actual-path> ::= <RRO>
```

In this standard, we extend these two elements to allow multiple ERO/RRO objects to be present in the <intended-path>/<actual-path>:

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

```
<actual-path> ::= (<RRO>|
                  (<PATH-ATTRIB><RRO>)
                  [<actual-path>])
```

7. Examples

7.1. SR Policy Candidate-Path with Multiple Segment-Lists

Consider the following sample SR Policy, taken from [I-D.ietf-spring-segment-routing-policy].

```
SR policy POL1 <headend, color, endpoint>
  Candidate-path CP1 <protocol-origin = 20, originator =
    100:1.1.1.1, discriminator = 1>
    Preference 200
    Weight W1, SID-List1 <SID1l...SID1i>
    Weight W2, SID-List2 <SID2l...SID2j>
  Candidate-path CP2 <protocol-origin = 20, originator =
    100:2.2.2.2, discriminator = 2>
    Preference 100
    Weight W3, SID-List3 <SID3l...SID3i>
    Weight W4, SID-List4 <SID4l...SID4j>
```

As specified in [I-D.ietf-pce-segment-routing-policy-cp], CP1 and CP2 are signaled as separate state-report elements and each has a unique PLSP-ID, assigned by the PCC. Let us assign PLSP-ID 100 to CP1 and PLSP-ID 200 to CP2.

The state-report for CP1 can be encoded as:

```

<state-report> =
  <LSP PLSP_ID=100>
  <ASSOCIATION>
  <END-POINT>
  <PATH-ATTRIB Path_ID=1 <WEIGHT-TLV Weight=W1>>
  <ERO SID-List1>
  <PATH-ATTRIB Path_ID=2 <WEIGHT-TLV Weight=W2>>
  <ERO SID-List2>

```

The state-report for CP2 can be encoded as:

```

<state-report> =
  <LSP PLSP_ID=200>
  <ASSOCIATION>
  <END-POINT>
  <PATH-ATTRIB Path_ID=1 <WEIGHT-TLV Weight=W3>>
  <ERO SID-List3>
  <PATH-ATTRIB Path_ID=2 <WEIGHT-TLV Weight=W4>>
  <ERO SID-List4>

```

The above sample state-report elements only specify the minimum mandatory objects, of course other objects like SRP, LSPA, METRIC, etc., are allowed to be inserted.

Note that the syntax

```
<PATH-ATTRIB Path_ID=1 <WEIGHT-TLV Weight=W1>>
```

, simply means that this is PATH-ATTRIB object with Path ID field set to "1" and with a MULTIPATH-WEIGHT TLV carrying weight of "W1".

7.2. Two Primary Paths Protected by One Backup Path

Suppose there are 3 paths: A, B, C. Where A,B are primary and C is to be used only when A or B fail. Suppose the Path IDs for A, B, C are respectively 1, 2, 3. This would be encoded in a state-report as:

```

<state-report> =
  <LSP>
  <ASSOCIATION>
  <END-POINT>
  <PATH-ATTRIB Path_ID=1 <BACKUP-TLV B=0, Backup_Paths=[3]>>
  <ERO A>
  <PATH-ATTRIB Path_ID=2 <BACKUP-TLV B=0, Backup_Paths=[3]>>
  <ERO B>
  <PATH-ATTRIB Path_ID=3 <BACKUP-TLV B=1, Backup_Paths=[]>>
  <ERO C>

```

Note that the syntax

```
<PATH-ATTRIB Path_ID=1 <BACKUP-TLV B=0, Backup_Paths=[3]>>
```

, simply means that this is PATH-ATTRIB object with Path ID field set to "1" and with a MULTIPATH-BACKUP TLV that has B-flag cleared and contains a single backup path with Backup Path ID of 3.

7.3. Composite Candidate Path

Consider the following Composite Candidate Path, taken from [I-D.ietf-spring-segment-routing-policy].

```
SR policy POL100 <headend = H1, color = 100, endpoint = E1>  
  Candidate-path CP1 <protocol-origin = 20, originator =  
    100:1.1.1.1, discriminator = 1>  
    Preference 200  
    Weight W1, SR policy <color = 1>  
    Weight W2, SR policy <color = 2>
```

This is signaled in PCEP as:

```
<LSP PLSP_ID=100>  
  <ASSOCIATION>  
  <END-POINT>  
  <PATH-ATTRIB Path_ID=1  
    <WEIGHT-TLV Weight=W1>  
    <COLOR-TLV Color=1>>  
  <ERO (empty)>  
  <PATH-ATTRIB Path_ID=2  
    <WEIGHT-TLV Weight=W2>  
    <COLOR-TLV Color=2>>  
  <ERO (empty)>
```

7.4. Opposite Direction Tunnels

Consider the two opposite-direction SR Policies between end-points H1 and E1.

```

SR policy POL1 <headend = H1, color, endpoint = E1>
  Candidate-path CP1
    Preference 200
    Bidirectional Association = A1
    SID-List = <H1,M1,M2,E1>
    SID-List = <H1,M3,M4,E1>
  Candidate-path CP2
    Preference 100
    Bidirectional Association = A2
    SID-List = <H1,M5,M6,E1>
    SID-List = <H1,M7,M8,E1>

SR policy POL2 <headend = E1, color, endpoint = H1>
  Candidate-path CP1
    Preference 200
    Bidirectional Association = A1
    SID-List = <E1,M2,M1,H1>
    SID-List = <E1,M4,M3,H1>
  Candidate-path CP2
    Preference 100
    Bidirectional Association = A2
    SID-List = <E1,M6,M5,H1>

```

The state-report for POL1, CP1 can be encoded as:

```

<state-report> =
  <LSP PLSP_ID=100>
  <BIDIRECTIONAL ASSOCIATION = A1>
  <PATH-ATTRIB PathID=1 R-flag=0
    <OPPDIR-PATH-TLV OppositePathID=3>>
  <ERO <H1,M1,M2,E1>>
  <PATH-ATTRIB PathID=2 R-flag=0
    <OPPDIR-PATH-TLV OppositePathID=4>>
  <ERO <H1,M3,M4,E1>>
  <PATH-ATTRIB PathID=3 R-flag=1
    <OPPDIR-PATH-TLV OppositePathID=1>>
  <ERO <E1,M2,M1,H1>>
  <PATH-ATTRIB PathID=4 R-flag=1
    <OPPDIR-PATH-TLV OppositePathID=2>>
  <ERO <E1,M4,M3,H1>>

```

The state-report for POL1, CP2 can be encoded as:

```

<state-report> =
  <LSP PLSP_ID=200>
  <BIDIRECTIONAL ASSOCIATION = A2>
  <PATH-ATTRIB PathID=1 R-flag=0
    <OPPDIR-PATH-TLV OppositePathID=3>>
  <ERO <H1,M5,N6,E1>>
  <PATH-ATTRIB PathID=2 R-flag=0
    <OPPDIR-PATH-TLV OppositePathID=0>>
  <ERO <H1,M7,M8,E1>>
  <PATH-ATTRIB PathID=3 R-flag=1
    <OPPDIR-PATH-TLV OppositePathID=1>>
  <ERO <E1,M6,M5,H1>>

```

The state-report for POL2, CP1 can be encoded as:

```

<state-report> =
  <LSP PLSP_ID=100>
  <BIDIRECTIONAL ASSOCIATION = A1>
  <PATH-ATTRIB PathID=1 R-flag=0
    <OPPDIR-PATH-TLV OppositePathID=3>>
  <ERO <E1,M2,M1,H1>>
  <PATH-ATTRIB PathID=2 R-flag=0
    <OPPDIR-PATH-TLV OppositePathID=4>>
  <ERO <E1,M4,M3,H1>>
  <PATH-ATTRIB PathID=3 R-flag=1
    <OPPDIR-PATH-TLV OppositePathID=1>>
  <ERO <H1,M1,M2,E1>>
  <PATH-ATTRIB PathID=4 R-flag=1
    <OPPDIR-PATH-TLV OppositePathID=2>>
  <ERO <H1,M3,M4,E1>>

```

The state-report for POL2, CP2 can be encoded as:

```

<state-report> =
  <LSP PLSP_ID=200>
  <BIDIRECTIONAL ASSOCIATION = A2>
  <PATH-ATTRIB PathID=1 R-flag=0
    <OPPDIR-PATH-TLV OppositePathID=3>>
  <ERO <E1,M6,M5,H1>>
  <PATH-ATTRIB PathID=2 R-flag=1
    <OPPDIR-PATH-TLV OppositePathID=0>>
  <ERO <H1,M7,M8,E1>>
  <PATH-ATTRIB PathID=3 R-flag=1
    <OPPDIR-PATH-TLV OppositePathID=1>>
  <ERO <H1,M5,N6,E1>>

```

8. Implementation Status

Note to the RFC Editor - remove this section before publication, as well as remove the reference to [RFC7942].

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".

8.1. Cisco Systems

Organization: Cisco Systems
Implementation: IOS-XR PCC and PCE
Description: Circuit-Style SR Policies
Maturity Level: Supported feature
Coverage: Multiple Segment-Lists and reverse paths in SR Policy
Contact: mkoldych@cisco.com

8.2. Ciena Corp

Organization: Ciena Corp
Implementation: Head-end and controller
Maturity Level: Proof of concept
Coverage: Full
Contact: byadav@ciena.com

9. IANA Considerations

9.1. PCEP Object

IANA is requested to make the assignment of a new value for the existing "PCEP Objects" registry as follows:

Object-Class Value	Name	Object-Type Value	Reference
TBD2	PATH-ATTRIB	1	This document

9.2. PCEP TLV

IANA is requested to make the assignment of a new value for the existing "PCEP TLV Type Indicators" registry as follows:

TLV Type Value	TLV Name	Reference
TBD1	MULTIPATH-CAP	This document
TBD3	MULTIPATH-WEIGHT	This document
TBD4	MULTIPATH-BACKUP	This document
TBD9	MULTIPATH-OPPDIR-PATH	This document

9.3. PCEP-Error Object

IANA is requested to make the assignment of a new value for the existing "PCEP-ERROR Object Error Types and Values" sub-registry of the PCEP Numbers registry for the following errors:

Error-Type	Error-Value	Reference
10	TBD5 - Conflicting Path ID	This document
19	TBD7 - Not supported path backup	This document
19	TBD8 - Non-empty path	This document

9.4. Flags in the Multipath Capability TLV

IANA is requested to create a new sub-registry to manage the Flag field of the MULTIPATH-CAP TLV, called "Flags in MULTIPATH-CAP TLV". New values are to be assigned by Standards Action [RFC8126]

Bit	Description	Reference
0-12	Unassigned	This document
13	O-flag: support for processing MULTIPATH-OPPDIR-PATH TLV	This document
14	B-flag: support for processing MULTIPATH-BACKUP TLV	This document
15	W-flag: support for processing MULTIPATH-WEIGHT TLV	This document

9.5. Flags in the Path Attribute Object

IANA is requested to create a new sub-registry to manage the Flag field of the PATH-ATTRIBUTE object, called "Flags in PATH-ATTRIBUTE Object". New values are to be assigned by Standards Action [RFC8126]

Bit	Description	Reference
0-12	Unassigned	This document
13-15	O-flag: Operational state	This document

9.6. Flags in the Multipath Backup TLV

IANA is requested to create a new sub-registry to manage the Flag field of the MULTIPATH-BACKUP TLV, called "Flags in MULTIPATH-BACKUP TLV". New values are to be assigned by Standards Action [RFC8126]

Bit	Description	Reference
0-14	Unassigned	This document
15	B-flag: Pure backup	This document

9.7. Flags in the Multipath Opposite Direction Path TLV

IANA is requested to create a new sub-registry to manage the flag fields of the MULTIPATH-OPPDIR-PATH TLV, called "Flags in the MULTIPATH-OPPDIR-PATH TLV". New values are to be assigned by Standards Action [RFC8126]

Bit	Description	Reference
0-12	Unassigned	This document
14	L-flag: Link co-routed	This document
15	N-flag: Node co-routed	This document

10. Security Considerations

None at this time.

11. Acknowledgement

Thanks to Dhruv Dhody for ideas and discussion.

12. Contributors

Andrew Stone
Nokia

Email: andrew.stone@nokia.com

13. References

13.1. Normative References

- [I-D.draft-rajagopalan-pce-pcep-color]
Rajagopalan, B., Beeram, V. P., Peng, S., Xiong, Q., Koldychev, M., and G. Mishra, "Path Computation Element Protocol (PCEP) Extension for Color", Work in Progress, Internet-Draft, draft-rajagopalan-pce-pcep-color-01, 14 November 2021, <<https://www.ietf.org/archive/id/draft-rajagopalan-pce-pcep-color-01.txt>>.
- [I-D.ietf-pce-segment-routing-policy-cpl]
Koldychev, M., Sivabalan, S., Barth, C., Peng, S., and H. Bidgoli, "PCEP extension to support Segment Routing Policy Candidate Paths", Work in Progress, Internet-Draft, draft-

ietf-pce-segment-routing-policy-cp-06, 22 October 2021,
<<https://www.ietf.org/archive/id/draft-ietf-pce-segment-routing-policy-cp-06.txt>>.

[I-D.ietf-spring-segment-routing-policy]
Filsfils, C., Talaulikar, K., Voyer, D., Bogdanov, A., and
P. Mattes, "Segment Routing Policy Architecture", Work in
Progress, Internet-Draft, draft-ietf-spring-segment-
routing-policy-22, 22 March 2022,
<<https://www.ietf.org/archive/id/draft-ietf-spring-segment-routing-policy-22.txt>>.

[I-D.koldychev-pce-operational]
Koldychev, M., Sivabalan, S., Peng, S., Achaval, D., and
H. Kotni, "PCEP Operational Clarification", Work in
Progress, Internet-Draft, draft-koldychev-pce-operational-
05, 19 February 2022, <<https://www.ietf.org/archive/id/draft-koldychev-pce-operational-05.txt>>.

[I-D.schmutzer-pce-cs-sr-policy]
Schmutzer, C., Filsfils, C., Ali, Z., Clad, F., and P.
Maheshwari, "Circuit Style Segment Routing Policies", Work
in Progress, Internet-Draft, draft-schmutzer-pce-cs-sr-
policy-01, 7 March 2022, <<https://www.ietf.org/archive/id/draft-schmutzer-pce-cs-sr-policy-01.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>>.

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

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

- [RFC8231] Crabbe, E., Minei, I., Medved, J., and R. Varga, "Path Computation Element Communication Protocol (PCEP) Extensions for Stateful PCE", RFC 8231, DOI 10.17487/RFC8231, September 2017, <<https://www.rfc-editor.org/info/rfc8231>>.
- [RFC8281] Crabbe, E., Minei, I., Sivabalan, S., and R. Varga, "Path Computation Element Communication Protocol (PCEP) Extensions for PCE-Initiated LSP Setup in a Stateful PCE Model", RFC 8281, DOI 10.17487/RFC8281, December 2017, <<https://www.rfc-editor.org/info/rfc8281>>.
- [RFC8664] Sivabalan, S., Filsfils, C., Tantsura, J., Henderickx, W., and J. Hardwick, "Path Computation Element Communication Protocol (PCEP) Extensions for Segment Routing", RFC 8664, DOI 10.17487/RFC8664, December 2019, <<https://www.rfc-editor.org/info/rfc8664>>.

13.2. Informative References

- [RFC4655] Farrel, A., Vasseur, J.-P., and J. Ash, "A Path Computation Element (PCE)-Based Architecture", RFC 4655, DOI 10.17487/RFC4655, August 2006, <<https://www.rfc-editor.org/info/rfc4655>>.
- [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>>.
- [RFC8745] Ananthakrishnan, H., Sivabalan, S., Barth, C., Minei, I., and M. Negi, "Path Computation Element Communication Protocol (PCEP) Extensions for Associating Working and Protection Label Switched Paths (LSPs) with Stateful PCE", RFC 8745, DOI 10.17487/RFC8745, March 2020, <<https://www.rfc-editor.org/info/rfc8745>>.

Authors' Addresses

Mike Koldychev
Cisco Systems, Inc.
Email: mkoldych@cisco.com

Siva Sivabalan
Ciena Corporation
Email: ssivabal@ciena.com

Tarek Saad
Juniper Networks, Inc.
Email: tsaad@juniper.net

Vishnu Pavan Beeram
Juniper Networks, Inc.
Email: vbeeram@juniper.net

Hooman Bidgoli
Nokia
Email: hooman.bidgoli@nokia.com

Bhupendra Yadav
Ciena
Email: byadav@ciena.com

Shuping Peng
Huawei Technologies
Email: pengshuping@huawei.com

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

PCE Working Group
Internet-Draft
Intended status: Standards Track
Expires: 25 April 2022

M. Koldychev
Cisco Systems, Inc.
S. Sivabalan
Ciena Corporation
C. Barth
Juniper Networks, Inc.
S. Peng
Huawei Technologies
H. Bidgoli
Nokia
October 2021

PCEP extension to support Segment Routing Policy Candidate Paths
draft-ietf-pce-segment-routing-policy-cp-06

Abstract

This document introduces a mechanism to specify a Segment Routing (SR) policy, as a collection of SR candidate paths. An SR policy is identified by <headend, color, endpoint> tuple. An SR policy can contain one or more candidate paths where each candidate path is identified in PCEP by its uniquely assigned PLSP-ID. This document proposes extension to PCEP to support association among candidate paths of a given SR policy. The mechanism proposed in this document is applicable to both MPLS and IPv6 data planes of SR.

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 4 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	4
3. Motivation	5
3.1. Group Candidate Paths belonging to the same SR policy . .	5
3.2. Instantiation of SR policy candidate paths	5
3.3. Avoid computing lower preference candidate paths	5
3.4. Minimal signaling overhead	6
4. Procedure	6
4.1. Overview	6
4.1.1. SR Policy Identifiers	7
4.1.2. SR Policy Candidate Path Identifiers	7
4.1.3. SR Policy Candidate Path Attributes	7
4.2. Multiple Optimization Objectives and Constraints	8
5. SR Policy Association	8
5.1. Association Parameters	8
5.2. Association Information	10
5.2.1. SR Policy Name TLV	10
5.2.2. SR Policy Candidate Path Identifiers TLV	11
5.2.3. SR Policy Candidate Path Name TLV	12
5.2.4. SR Policy Candidate Path Preference TLV	12
6. Generic Mechanisms	13
6.1. Computation Priority TLV	13
6.2. Explicit Null Label Policy (ENLP) TLV	13
6.3. Invalidation TLV	14
6.4. Specified-BSID-only	15

7. Examples	15
7.1. PCC Initiated SR Policy with single candidate-path . . .	15
7.2. PCC Initiated SR Policy with multiple candidate-paths . .	16
7.3. PCE Initiated SR Policy with single candidate-path . . .	16
7.4. PCE Initiated SR Policy with multiple candidate-paths . .	17
8. IANA Considerations	17
8.1. Association Type	17
8.2. PCEP TLV Type Indicators	18
8.3. PCEP Errors	18
8.4. TE-PATH-BINDING TLV Flag field	19
9. Implementation Status	19
9.1. Cisco	20
9.2. Juniper	20
10. Security Considerations	20
11. Acknowledgement	21
12. References	21
12.1. Normative References	21
12.2. Informative References	22
Appendix A. Contributors	23
Authors' Addresses	24

1. Introduction

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

PCEP Extensions for the Stateful PCE Model [RFC8231] describes a set of extensions to PCEP to enable active control of Multiprotocol Label Switching Traffic Engineering (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.

PCEP Extensions for Segment Routing [RFC8664] specifies extensions to the Path Computation Element Protocol (PCEP) that allow a stateful PCE to compute and initiate Traffic Engineering (TE) paths, as well as a PCC to request a path subject to certain constraint(s) and optimization criteria in SR networks.

PCEP Extensions for Establishing Relationships Between Sets of LSPs [RFC8697] introduces a generic mechanism to create a grouping of LSPs which can then be used to define associations between a set of LSPs and a set of attributes (such as configuration parameters or behaviors) and is equally applicable to stateful PCE (active and passive modes) and stateless PCE.

Segment Routing Policy for Traffic Engineering
[I-D.ietf-spring-segment-routing-policy] details the concepts of SR Policy and approaches to steering traffic into an SR Policy.

An SR Policy contains one or more SR Policy Candidate Paths where one or more such paths can be computed via PCE. This document specifies PCEP extensions to signal additional information to map candidate paths to their SR policies. Each candidate path maps to a unique PLSP-ID in PCEP. By associating multiple candidate paths together, a PCE becomes aware of the hierarchical structure of an SR policy. Thus the PCE can take computation and control decisions about the candidate paths, with the additional knowledge that these candidate paths belong to the same SR policy. This is accomplished via the use of the existing PCEP Association object, by defining a new association type specifically for associating SR candidate paths into a single SR policy.

2. Terminology

The following terminologies are used in this document:

Endpoint: The IPv4 or IPv6 endpoint address of the SR policy in question, as described in [I-D.ietf-spring-segment-routing-policy].

Association Parameters: As described in [RFC8697], the combination of the mandatory fields Association Type, Association ID and Association Source in the ASSOCIATION object uniquely identify the association group. If the optional TLVs - Global Association Source or Extended Association ID are included, then they MUST be included in combination with mandatory fields to uniquely identify the association group.

Association Information: As described in [RFC8697], the ASSOCIATION object could also include other TLVs based on the association types, that provides non-key information.

SRPAG: SR Policy Association Group.

SRPAT: SR Policy Association Type.

SRPAT ASSOCIATION: ASSOCIATION object of type SR Policy Association.

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

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

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

PCEP: Path Computation Element Protocol.

PCEP Tunnel: The entity identified by the PLSP-ID, as per [I-D.koldychev-pce-operational].

3. Motivation

The SR Policy Association and its TLVs, defined in this document, allow PCEP speakers to exchange additional information about SR Policy Candidate Paths and their container SR Policy.

3.1. Group Candidate Paths belonging to the same SR policy

Since each SR Policy Candidate Path appears as a different Tunnel (identified via a PLSP-ID) in PCEP, it is useful to group together all the SR Policy Candidate Paths that belong to the same SR Policy. Furthermore, it is useful for the PCE to have knowledge of the SR Policy related information such as color, endpoint, protocol origin, discriminator, and preference.

3.2. Instantiation of SR policy candidate paths

A PCE needs to instantiate one or more SR Policy Candidate Paths on the PCC, as specified in [RFC8281]. Each SR Policy Candidate Path is identified by the tuple <headend, color, endpoint, originator, discriminator, preference>. This draft provides a mechanism to signal this information in PCEP.

3.3. Avoid computing lower preference candidate paths

When a PCE knows that a given set of SR Policy Candidate Paths all belong to the same SR Policy, then path computation MAY be done on only the highest preference candidate-path(s). Path computation for lower preference paths is not necessary if one or two higher preference paths are already computed. Since computing their paths will not affect traffic steering, it MAY be postponed until the higher preference paths become invalid.

3.4. Minimal signaling overhead

When an SR Policy contains multiple SR Policy Candidate Paths computed by a PCE, such candidate paths can be created, updated and deleted independently of each other. This is achieved by making each SR Policy Candidate Path correspond to a unique Tunnel (identified via PLSP-ID). For example, if an SR Policy has 4 SR Policy Candidate Paths, then if the PCE wants to update one of those, only one set of PCUpd and PCRpt messages needs to be exchanged.

4. Procedure

4.1. Overview

As per [RFC8697], LSPs are placed into an association group. As per [I-D.koldychev-pce-operational], LSPs are contained in PCEP Tunnels and a PCEP Tunnel is contained in an Association if all of its LSPs are in that Association. PCEP Tunnels naturally map to SR Policy Candidate Paths and PCEP Associations naturally map to SR Policies.

The mapping between PCEP Associations and SR Policies is always one-to-one. However, the mapping between PCEP Tunnels and SR Policy Candidate Paths may be either one-to-one, or many-to-one, see Section 4.2.

Each SR Policy Candidate Path contains one or more Segment Lists. The subject of encoding multiple Segment Lists within an SR Policy Candidate Path is described in [I-D.koldychev-pce-multipath].

This document defines a new Association Type called "SR Policy Association", of value 6 based on the generic ASSOCIATION object. The new Association Type is also called "SRPAT", for "SR Policy Association Type". We say "SRPAT ASSOCIATION" to mean "ASSOCIATION object of type SR Policy Association". The group of LSPs that are part of the SR Policy Association is called "SRPAG", for "SR Policy Association Group".

As per the processing rules specified in section 6.4 of [RFC8697], if a PCEP speaker does not support the SRPAT, it MUST return a PCErr message with Error-Type = 26 "Association Error", Error-Value = 1 "Association-type is not supported".

A given LSP MUST belong to at most one SRPAG, since an SR Policy Candidate Path cannot belong to multiple SR Policies. If a PCEP speaker receives a PCEP message with more than one SRPAT ASSOCIATION for the same LSP, then the PCEP speaker MUST send a PCErr message with Error-Type = 26 "Association Error", Error-Value = 7 "Cannot join the association group".

An SRPAT ASSOCIATION carries three pieces of information: SR Policy Identifiers, SR Policy Candidate Path Identifiers, and SR Policy Candidate Path Attributes.

4.1.1. SR Policy Identifiers

SR Policy Identifiers uniquely identify the SR policy within the context of the headend. SR Policy Identifiers MUST be the same for all SR Policy Candidate Paths in the same SRPAG. SR Policy Identifiers MUST NOT change for a given SR Policy Candidate Path during its lifetime. SR Policy Identifiers MUST be different for different SRPAGs. SR Policy Identifiers consist of:

- * Headend router where the SR Policy originates.
- * Color of SR Policy.
- * Endpoint of SR Policy.

4.1.2. SR Policy Candidate Path Identifiers

SR Policy Candidate Path Identifiers uniquely identify the SR Policy Candidate Path within the context of an SR Policy. SR Policy Candidate Path Identifiers MUST NOT change for a given LSP during its lifetime. SR Policy Candidate Path Identifiers MUST be different for different LSPs within the same SRPAG. When these rules are not satisfied, the PCE MUST send a PCErr message with Error-Type = 26 "Association Error", Error Value = TBD8 "SR Policy Candidate Path Identifiers Mismatch". SR Policy Candidate Path Identifiers consist of:

- * Protocol Origin.
- * Originator.
- * Discriminator.

4.1.3. SR Policy Candidate Path Attributes

SR Policy Candidate Path Attributes carry non-key information about the candidate path and MAY change during the lifetime of the LSP. SR Policy Candidate Path Attributes consist of:

- * Preference.
- * Optionally, the SR Policy Candidate Path name.
- * Optionally, the SR Policy name.

4.2. Multiple Optimization Objectives and Constraints

In certain scenarios, it is desired for each SR Policy Candidate Path to contain multiple sub-candidate paths, each of which has a different optimization objective and constraints. Traffic is then sent ECMP or UCMP among these sub-candidate paths.

This is represented in PCEP by a many-to-one mapping between PCEP Tunnels and SR Policy Candidate Paths. This means that multiple PCEP Tunnels are allocated for each SR Policy Candidate Path. Each PCEP Tunnel has its own optimization objective and constraints. When a single SR Policy Candidate Path contains multiple PCEP Tunnels, each of these PCEP Tunnels MUST have identical values of Candidate Path Identifiers, as encoded in SRPOLICY-CPATH-ID TLV, see Section 5.2.2.

5. SR Policy Association

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 (6) "SR Policy Association". This Association Type is dynamic in nature, thus operator-configured Association Range MUST NOT be set for this Association type and MUST be ignored.

5.1. Association Parameters

As per [I-D.ietf-spring-segment-routing-policy], an SR Policy is identified through the tuple <headend, color, endpoint>. the headend is encoded as the Association Source in the ASSOCIATION object and the color and endpoint are encoded as part of Extended Association ID TLV.

The Association Parameters (see Section 2) consist of:

- * Association Type: set to 6 "SR Policy Association".
- * Association Source (IPv4/IPv6): set to the headend IP address.
- * Association ID (16-bit): set to "1".
- * Extended Association ID TLV: encodes the Color and Endpoint of the SR Policy.

The Association Source MUST be set to the headend value of the SR Policy, as defined in [I-D.ietf-spring-segment-routing-policy] Section 2.1. If the PCC receives a PCInit message for a non-existent SR Policy, where the Association Source is set not to the headend value but to some globally unique IP address that the PCC owns, then

the PCC SHOULD accept the PCInit message and create the SR Policy Association with the Association Source that was sent in the PCInit message.

The 16-bit Association ID field in the ASSOCIATION object MUST be set to the value of "1".

The Extended Association ID TLV MUST be included and it MUST be in the following format:

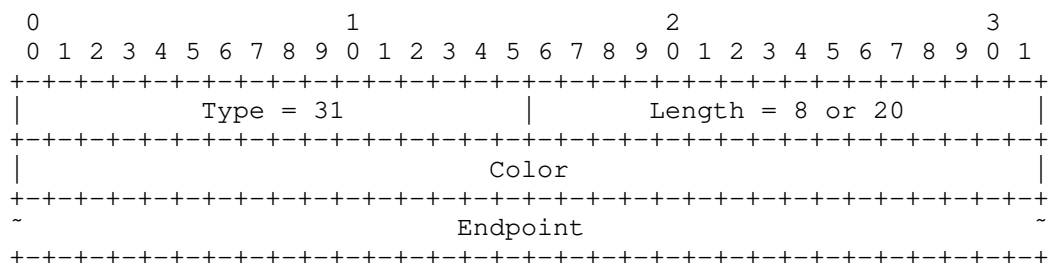


Figure 1: Extended Association ID TLV format

Type: Extended Association ID TLV, type = 31.

Length: Either 8 or 20, depending on whether IPv4 or IPv6 address is encoded in the Endpoint.

Color: SR Policy color value.

Endpoint: can be either IPv4 or IPv6, depending on whether the policy endpoint is IPv4 or IPv6. This value MAY be different from the one contained in the END-POINTS object, or in the LSP IDENTIFIERS TLV of the LSP object. This value is part of the tuple <color, endpoint> that identifies the SR Policy on a given headend.

If the PCEP speaker receives an SRPAT ASSOCIATION whose Association Parameters do not follow the above specification, then the PCEP speaker MUST send PCErr message with Error-Type = 26 "Association Error", Error-Value = TBD7 "SR Policy Identifiers Mismatch".

The purpose of choosing the Association Parameters in this way is to guarantee that there is no possibility of a race condition when multiple PCEP speakers want to create the same SR Policy at the same time. By adhering to this format, all PCEP speakers come up with the same Association Parameters independently of each other. Thus, there is no chance that different PCEP speakers will come up with different Association Parameters for the same SR Policy.

5.2. Association Information

The SRPAT ASSOCIATION contains the following TLVs:

- * SRPOLICY-POL-NAME TLV: (optional) encodes SR Policy Name string.
- * SRPOLICY-CPATH-ID TLV: (mandatory) encodes SR Policy Candidate Path Identifiers.
- * SRPOLICY-CPATH-NAME TLV: (optional) encodes SR Policy Candidate Path string name.
- * SRPOLICY-CPATH-PREFERENCE TLV: (optional) encodes SR Policy Candidate Path preference value.

Of these new TLVs, SRPOLICY-CPATH-ID TLV is mandatory. When a mandatory TLV is missing from the SRPAT ASSOCIATION object, the PCE MUST send a PCErr message with Error-Type = 6 "Mandatory Object Missing", Error-Value = TBD6 "Missing Mandatory TLV".

5.2.1. SR Policy Name TLV

The SRPOLICY-POL-NAME TLV is an optional TLV for the SRPAT ASSOCIATION. At most one SRPOLICY-POL-NAME TLV SHOULD be encoded by the sender and only the first occurrence is processed and any others MUST be ignored.

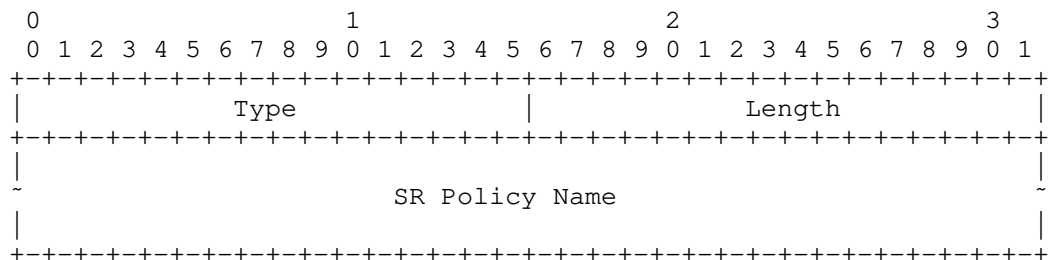


Figure 2: The SRPOLICY-POL-NAME TLV format

Type: 56 for "SRPOLICY-POL-NAME" TLV.

Length: indicates the length of the value portion of the TLV in octets and MUST be greater than 0. The TLV MUST be zero-padded so that the TLV is 4-octet aligned.

SR Policy Name: SR Policy name, as defined in [I-D.ietf-spring-segment-routing-policy]. It SHOULD be a string of printable ASCII characters, without a NULL terminator.

5.2.2. SR Policy Candidate Path Identifiers TLV

The SRPOLICY-CPATH-ID TLV is a mandatory TLV for the SRPAT ASSOCIATION. Only one SRPOLICY-CPATH-ID TLV SHOULD be encoded by the sender and only the first occurrence is processed and any others MUST be ignored.

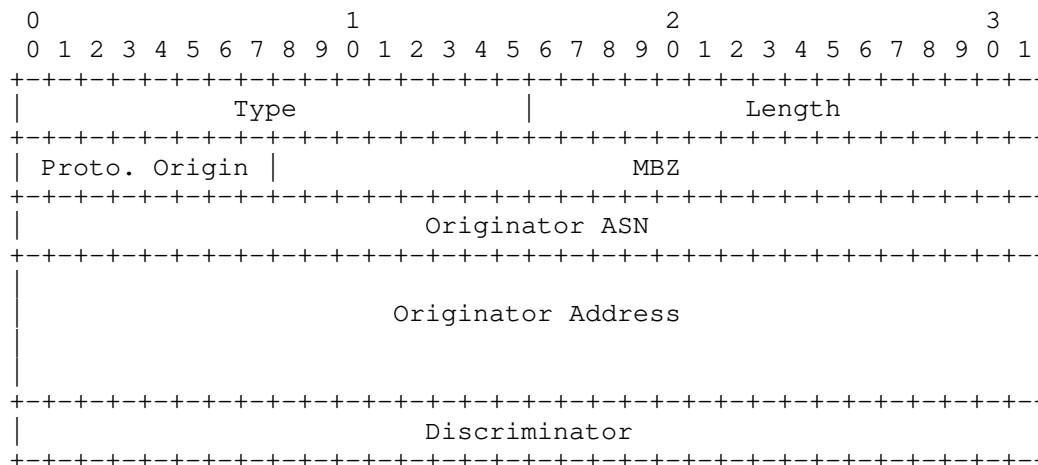


Figure 3: The SRPOLICY-CPATH-ID TLV format

Type: 57 for "SRPOLICY-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. Note that in PCInit messages, the Protocol Origin is always set to "PCEP".

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

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

5.2.3. SR Policy Candidate Path Name TLV

The SRPOLICY-CPATH-NAME TLV is an optional TLV for the SRPAT ASSOCIATION. At most one SRPOLICY-CPATH-NAME TLV SHOULD be encoded by the sender and only the first occurrence is processed and any others MUST be ignored.

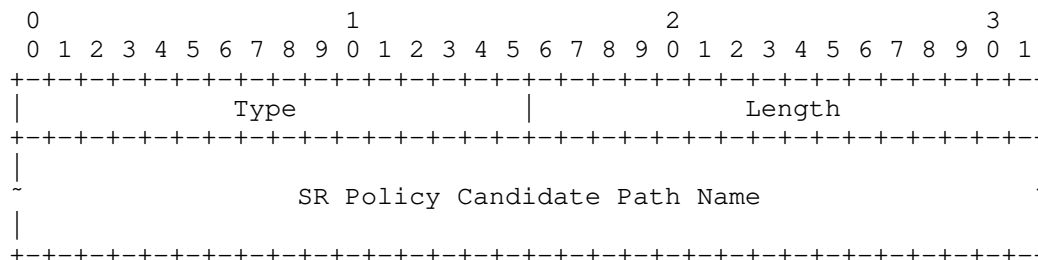


Figure 4: The SRPOLICY-CPATH-NAME TLV format

Type: 58 for "SRPOLICY-CPATH-NAME" TLV.

Length: indicates the length of the value portion of the TLV in octets and MUST be greater than 0. The TLV MUST be zero-padded so that the TLV is 4-octet aligned.

SR Policy Candidate Path Name: SR Policy Candidate Path Name, as defined in [I-D.ietf-spring-segment-routing-policy]. It SHOULD be a string of printable ASCII characters, without a NULL terminator.

5.2.4. SR Policy Candidate Path Preference TLV

The SRPOLICY-CPATH-PREFERENCE TLV is an optional TLV for the SRPAT ASSOCIATION. Only one SRPOLICY-CPATH-PREFERENCE TLV SHOULD be encoded by the sender and only the first occurrence is processed and any others MUST be ignored.

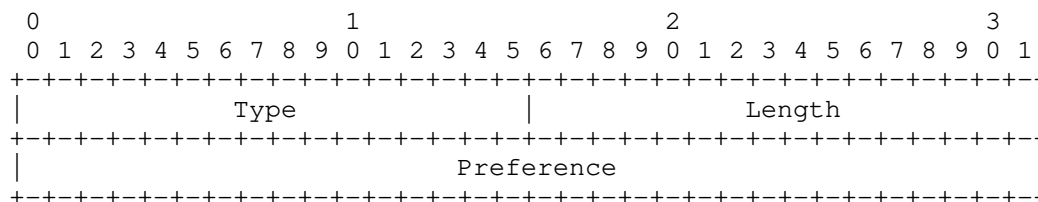


Figure 5: The SRPOLICY-CPATH-PREFERENCE TLV format

Type: 59 for "SRPOLICY-CPATH-PREFERENCE" TLV.

Length: 4.

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

If the TLV is missing, a default Preference value of 100 is used, as specified in Section 2.7 of [I-D.ietf-spring-segment-routing-policy].

6. Generic Mechanisms

This section describes various mechanisms that are standardized for SR Policies in [I-D.ietf-spring-segment-routing-policy], but are equally applicable to other tunnel types, such as RSVP-TE tunnels. Hence this section does not make use of the SRPAT ASSOCIATION.

6.1. Computation Priority TLV

The COMPUTATION-PRIORITY TLV is an optional TLV for the LSP object. It is used to signal the numerical computation priority, as specified in Section 2.12 of [I-D.ietf-spring-segment-routing-policy]. If the TLV is absent from the LSP object, a default Priority value of 128 is used.

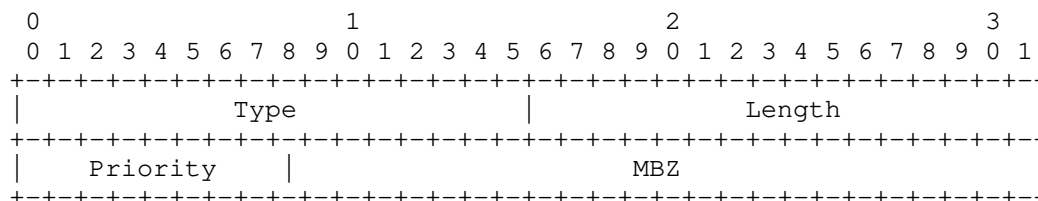


Figure 6: The COMPUTATION-PRIORITY TLV format

Type: TBD1 for "COMPUTATION-PRIORITY" TLV.

Length: 4.

Priority: Numerical priority with which this LSP is to be recomputed by the PCE upon topology change.

6.2. Explicit Null Label Policy (ENLP) TLV

The ENLP TLV is an optional TLV for the LSP object. It is used to implement the "Explicit Null Label Policy", as specified in Section 2.4.5 of [I-D.ietf-idr-segment-routing-te-policy].

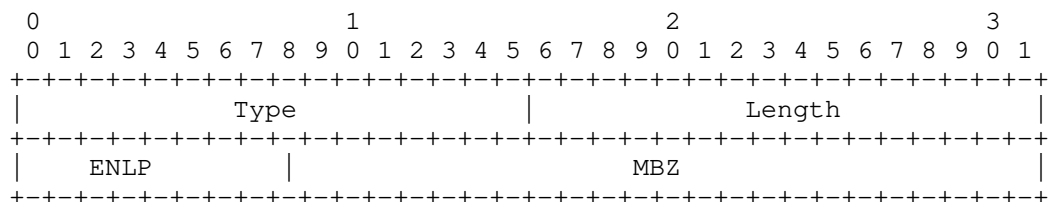


Figure 7: The Explicit Null Label Policy (ENLP) TLV format

Type: TBD2 for "ENLP" TLV.

Length: 4.

ENLP (Explicit NULL Label Policy): same values as in Section 2.4.5 of [I-D.ietf-idr-segment-routing-te-policy].

6.3. Invalidation TLV

The INVALIDATION TLV is an optional TLV for the LSP object. It is used to specify LSP behavior when the LSP is operationally down, in particular to facilitate the "Drop upon invalid" behavior, specified in Section 8.2 of [I-D.ietf-spring-segment-routing-policy].

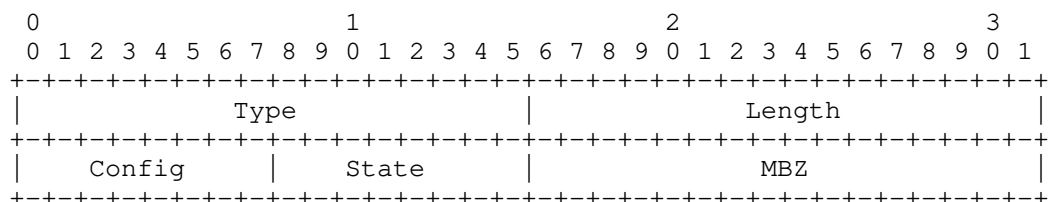


Figure 8: The INVALIDATION TLV format

Type: TBD3 for "INVALIDATION" TLV.

Length: 4.

Config: specifies the action to take when the LSP becomes invalid:

- * 0: (default) bring down the LSP and forward traffic somewhere else (i.e., IGP, etc.).
- * 1: drop traffic when the LSP is invalid.
- * 2-255: Reserved.

State: specifies the current state of the LSP:

- * 0: (default) traffic is not being dropped.
- * 1: traffic is being dropped, due to LSP being down and "Drop upon invalid" being set.
- * 2-255: Reserved.

The "State" field only has meaning when sent from PCC to the PCE in PCRpt messages, it is set to 0 when sent from PCE to PCC. The "Config" field is valid in both directions on the PCEP session, i.e., from PCC in PCRpt and from PCE in PCUpd and PCInit messages.

6.4. Specified-BSID-only

Specified-BSID-only functionality is defined in Section 6.2.3 of [I-D.ietf-spring-segment-routing-policy]. When specified-BSID-only is enabled for a particular binding SID, it means that the given binding SID is required to be allocated and programmed for the LSP to be operationally up. If the binding SID cannot be allocated or programmed for some reason, then the LSP must stay down.

To signal specified-BSID-only, a new bit: S (Specified-BSID-only) is allocated in the "TE-PATH-BINDING TLV Flag field" of the TE-PATH-BINDING TLV. When this bit is set for a particular BSID, it means that the BSID follows the Specified-BSID-only behavior. It is possible to have a mix of BSIDs for the same LSP: some with S=1 and some with S=0.

7. Examples

7.1. PCC Initiated SR Policy with single candidate-path

PCReq and PCRep messages are exchanged in the following sequence:

1. PCC sends PCReq message to the PCE, encoding the SRPAT ASSOCIATION and TLVs in the PCReq message.
2. PCE returns the path in PCRep message, and echoes back the SRPAT ASSOCIATION.

PCRpt and PCUpd messages are exchanged in the following sequence:

1. PCC sends PCRpt message to the PCE, including the LSP object and the SRPAT ASSOCIATION.
2. PCE computes path, possibly making use of the Association Information from the SRPAT ASSOCIATION.

3. PCE updates the SR policy candidate path's ERO using PCUpd message.

7.2. PCC Initiated SR Policy with multiple candidate-paths

PCRpt and PCUpd messages are exchanged in the following sequence:

1. For each candidate path of the SR Policy, the PCC generates a different PLSP-ID and symbolic-name and sends multiple PCRpt messages (or one message with multiple LSP objects) to the PCE. Each LSP object is followed by SRPAT ASSOCIATION with identical Color and Endpoint values. The Association Source is set to the IP address of the PCC and the Association ID is set to a number that PCC locally chose to represent the SR Policy.
2. PCE takes into account that all the LSPs belong to the same SR policy. PCE prioritizes computation for the highest preference LSP and sends PCUpd message(s) back to the PCC.
3. If a new candidate path is added on the PCC by the operator, then a new PLSP-ID and symbolic name is generated for that candidate path and a new PCRpt is sent to the PCE.
4. If an existing candidate path is removed from the PCC by the operator, then that PLSP-ID is deleted from the PCE by sending PCRpt with the R-flag in the LSP object set.

7.3. PCE Initiated SR Policy with single candidate-path

A candidate-path is created using the following steps:

1. PCE sends PCInitiate message, containing the SRPAT ASSOCIATION. The Association Source and the Association ID are set as described in Section 5.1.
2. PCC uses the color, endpoint and preference from the SRPAT ASSOCIATION to create a new candidate path. If no SR policy exists to hold the candidate path, then a new SR policy is created to hold the new candidate-path. The Originator of the candidate path is set to be the address of the PCE that is sending the PCInitiate message.
3. PCC sends a PCRpt message back to the PCE to report the newly created Candidate Path. The PCRpt message contains the SRPAT ASSOCIATION.

A candidate-path is deleted using the following steps:

1. PCE sends PCInitiate message, setting the R-flag in the LSP object.
2. PCC uses the PLSP-ID from the LSP object to find the candidate path and delete it. If this is the last candidate path under the SR policy, then the containing SR policy is deleted as well.

7.4. PCE Initiated SR Policy with multiple candidate-paths

A candidate-path is created using the following steps:

1. PCE sends a separate PCInitiate message for every candidate path that it wants to create, or it sends multiple LSP objects within a single PCInitiate message. The SRPAT ASSOCIATION is sent for every LSP in the PCInitiate message. The Association Source and the Association ID are set as described in Section 5.1.
2. PCC creates multiple candidate paths under the same SR policy, identified by Color and Endpoint.
3. PCC sends a PCRpt message back to the PCE to report the newly created Candidate Path. The PCRpt message contains the SRPAT ASSOCIATION. The Association Source and the Association ID are set as described in Section 5.1.

A candidate path is deleted using the following steps:

1. PCE sends PCInitiate message, setting the R-flag in the LSP object.
2. PCC uses the PLSP-ID from the LSP object to find the candidate path and delete it.

8. IANA Considerations

8.1. Association Type

This document defines a new association type: SR Policy Association. IANA is requested to make the following codepoint assignment in the "ASSOCIATION Type Field" subregistry [RFC8697] within the "Path Computation Element Protocol (PCEP) Numbers" registry:

Type	Name	Reference
6	SR Policy Association	This.I-D

8.2. PCEP TLV Type Indicators

This document defines four new TLVs for carrying additional information about SR policy and SR candidate paths. IANA is requested to make the assignment of a new value for the existing "PCEP TLV Type Indicators" registry as follows:

Value	Description	Reference
56	SRPOLICY-POL-NAME	This.I-D
57	SRPOLICY-CPATH-ID	This.I-D
58	SRPOLICY-CPATH-NAME	This.I-D
59	SRPOLICY-CPATH-PREFERENCE	This.I-D
TBD1	COMPUTATION-PRIORITY	This.I-D
TBD2	EXPLICIT-NULL-LABEL-POLICY	This.I-D
TBD3	INVALIDATION	This.I-D

8.3. PCEP Errors

This document defines one new Error-Value within the "Mandatory Object Missing" Error-Type and two new Error-Values within the "Association Error" Error-Type. IANA is requested to allocate new error values within the "PCEP-ERROR Object Error Types and Values" sub-registry of the PCEP Numbers registry, as follows:

Error-Type	Meaning	Error-value	Reference
6	Mandatory Object Missing		[RFC5440]
		TBD6: SR Policy Missing Mandatory TLV	This.I-D
26	Association Error		[RFC8697]
		TBD7: SR Policy Identifiers Mismatch	This.I-D
		TBD8: SR Policy Candidate Path Identifiers Mismatch	This.I-D

8.4. TE-PATH-BINDING TLV Flag field

IANA is requested to allocate new bit within the "TE-PATH-BINDING TLV Flag field" sub-registry of the PCEP Numbers registry, as follows:

Bit position	Description	Reference
1	Specified-BSID-only	This.I-D

9. 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".

9.1. Cisco

- * Organization: Cisco Systems
- * Implementation: IOS-XR PCC and PCE.
- * Description: An experimental code-point is currently used.
- * Maturity Level: Proof of concept.
- * Coverage: Full.
- * Contact: mkoldych@cisco.com

9.2. Juniper

- * Organization: Juniper Networks
- * Implementation: Head-end and controller.
- * Description: An experimental code-point is currently used.
- * Maturity Level: Proof of concept.
- * Coverage: Partial.
- * Contact: cbarth@juniper.net

10. Security Considerations

This document defines one new type for association, which do not add any new security concerns beyond those discussed in [RFC5440], [RFC8231], [RFC8664], [I-D.ietf-pce-segment-routing-ipv6] and [RFC8697] in itself.

The information carried in the SRPAT ASSOCIATION, as per this document is related to SR Policy. It often reflects information that can also be derived from the SR Database, but association provides a much easier grouping of related LSPs and messages. The SRPAT ASSOCIATION could provide an adversary with the opportunity to eavesdrop on the relationship between the LSPs. Thus securing the

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

11. Acknowledgement

Would like to thank Stephane Litkowski, Praveen Kumar and Tom Petch for review comments.

12. References

12.1. Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <<https://www.rfc-editor.org/info/rfc2119>>.
- [RFC5440] Vasseur, JP., Ed. and JL. Le Roux, Ed., "Path Computation Element (PCE) Communication Protocol (PCEP)", RFC 5440, DOI 10.17487/RFC5440, March 2009, <<https://www.rfc-editor.org/info/rfc5440>>.
- [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>>.
- [RFC7942] Sheffer, Y. and A. Farrel, "Improving Awareness of Running Code: The Implementation Status Section", BCP 205, RFC 7942, DOI 10.17487/RFC7942, July 2016, <<https://www.rfc-editor.org/info/rfc7942>>.
- [I-D.ietf-spring-segment-routing-policy] Filsfils, C., Talaulikar, K., Voyer, D., Bogdanov, A., and P. Mattes, "Segment Routing Policy Architecture", Work in Progress, Internet-Draft, draft-ietf-spring-segment-

routing-policy-13, 28 May 2021,
<<https://www.ietf.org/archive/id/draft-ietf-spring-segment-routing-policy-13.txt>>.

[I-D.ietf-idr-segment-routing-te-policy]

Previdi, S., Filsfils, C., Talaulikar, K., Mattes, P., Rosen, E., Jain, D., and S. Lin, "Advertising Segment Routing Policies in BGP", Work in Progress, Internet-Draft, draft-ietf-idr-segment-routing-te-policy-13, 7 June 2021, <<https://www.ietf.org/archive/id/draft-ietf-idr-segment-routing-te-policy-13.txt>>.

- [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>>.

- [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>>.

[I-D.koldychev-pce-operational]

Koldychev, M., Sivabalan, S., Peng, S., Achaval, D., and H. Kotni, "PCEP Operational Clarification", Work in Progress, Internet-Draft, draft-koldychev-pce-operational-04, 19 August 2021, <<https://www.ietf.org/archive/id/draft-koldychev-pce-operational-04.txt>>.

[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", Work in Progress, Internet-Draft, draft-koldychev-pce-multipath-05, 16 February 2021, <<https://www.ietf.org/archive/id/draft-koldychev-pce-multipath-05.txt>>.

12.2. Informative References

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

- [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>>.
- [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>>.
- [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", 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>>.

Appendix A. Contributors

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

Email: dhruv.ietf@gmail.com

Cheng Li
Huawei Technologies
Huawei Campus, No. 156 Beiqing Rd.
Beijing, 10095
China

Email: chengli13@huawei.com

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

Email: ssidor@cisco.com

Authors' Addresses

Mike Koldychev
Cisco Systems, Inc.
2000 Innovation Drive
Kanata Ontario K2K 3E8
Canada

Email: mkoldych@cisco.com

Siva Sivabalan
Ciena Corporation
385 Terry Fox Dr.
Kanata Ontario K2K 0L1
Canada

Email: ssivabal@ciena.com

Colby Barth
Juniper Networks, Inc.

Email: cbarth@juniper.net

Shuping Peng
Huawei Technologies
Huawei Campus, No. 156 Beiqing Rd.
Beijing
100095
China

Email: pengshuping@huawei.com

Hooman Bidgoli
Nokia

Email: hooman.bidgoli@nokia.com

PCE Working Group
Internet-Draft
Intended status: Standards Track
Expires: 21 October 2022

M. Koldychev
Cisco Systems, Inc.
S. Sivabalan
Ciena Corporation
C. Barth
Juniper Networks, Inc.
S. Peng
Huawei Technologies
H. Bidgoli
Nokia
April 2022

PCEP extension to support Segment Routing Policy Candidate Paths
draft-ietf-pce-segment-routing-policy-cp-07

Abstract

This document introduces a mechanism to specify a Segment Routing (SR) policy, as a collection of SR candidate paths. An SR policy is identified by <headend, color, endpoint> tuple. An SR policy can contain one or more candidate paths where each candidate path is identified in PCEP by its uniquely assigned PLSP-ID. This document proposes extension to PCEP to support association among candidate paths of a given SR policy. The mechanism proposed in this document is applicable to both MPLS and IPv6 data planes of SR.

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 3 October 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. Terminology	4
3. Motivation	5
3.1. Group Candidate Paths belonging to the same SR policy . .	5
3.2. Instantiation of SR policy candidate paths	5
3.3. Avoid computing lower preference candidate paths	5
3.4. Minimal signaling overhead	6
4. Procedure	6
4.1. Overview	6
4.1.1. SR Policy Identifiers	7
4.1.2. SR Policy Candidate Path Identifiers	7
4.1.3. SR Policy Candidate Path Attributes	7
4.2. Multiple Optimization Objectives and Constraints	8
5. SR Policy Association	8
5.1. Association Parameters	8
5.2. Association Information	10
5.2.1. SR Policy Name TLV	10
5.2.2. SR Policy Candidate Path Identifiers TLV	11
5.2.3. SR Policy Candidate Path Name TLV	12
5.2.4. SR Policy Candidate Path Preference TLV	12
6. Generic Mechanisms	13
6.1. Computation Priority TLV	13
6.2. Explicit Null Label Policy (ENLP) TLV	13
6.3. Invalidation TLV	14
6.4. Specified-BSID-only	15

7. Examples	16
7.1. PCC Initiated SR Policy with single candidate-path . . .	16
7.2. PCC Initiated SR Policy with multiple candidate-paths . .	16
7.3. PCE Initiated SR Policy with single candidate-path . . .	17
7.4. PCE Initiated SR Policy with multiple candidate-paths . .	17
8. IANA Considerations	18
8.1. Association Type	18
8.2. PCEP TLV Type Indicators	18
8.3. PCEP Errors	19
8.4. TE-PATH-BINDING TLV Flag field	19
9. Implementation Status	19
9.1. Cisco	20
9.2. Juniper	20
10. Security Considerations	21
11. Acknowledgement	21
12. References	21
12.1. Normative References	21
12.2. Informative References	23
Appendix A. Contributors	23
Authors' Addresses	24

1. Introduction

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

PCEP Extensions for the Stateful PCE Model [RFC8231] describes a set of extensions to PCEP to enable active control of Multiprotocol Label Switching Traffic Engineering (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.

PCEP Extensions for Segment Routing [RFC8664] specifies extensions to the Path Computation Element Protocol (PCEP) that allow a stateful PCE to compute and initiate Traffic Engineering (TE) paths, as well as a PCC to request a path subject to certain constraint(s) and optimization criteria in SR networks.

PCEP Extensions for Establishing Relationships Between Sets of LSPs [RFC8697] introduces a generic mechanism to create a grouping of LSPs which can then be used to define associations between a set of LSPs and a set of attributes (such as configuration parameters or behaviors) and is equally applicable to stateful PCE (active and passive modes) and stateless PCE.

Segment Routing Policy for Traffic Engineering
[I-D.ietf-spring-segment-routing-policy] details the concepts of SR Policy and approaches to steering traffic into an SR Policy.

An SR Policy contains one or more SR Policy Candidate Paths where one or more such paths can be computed via PCE. This document specifies PCEP extensions to signal additional information to map candidate paths to their SR policies. Each candidate path maps to a unique PLSP-ID in PCEP. By associating multiple candidate paths together, a PCE becomes aware of the hierarchical structure of an SR policy. Thus the PCE can take computation and control decisions about the candidate paths, with the additional knowledge that these candidate paths belong to the same SR policy. This is accomplished via the use of the existing PCEP Association object, by defining a new association type specifically for associating SR candidate paths into a single SR policy.

2. Terminology

The following terminologies are used in this document:

Endpoint: The IPv4 or IPv6 endpoint address of the SR policy in question, as described in [I-D.ietf-spring-segment-routing-policy].

Association Parameters: As described in [RFC8697], the combination of the mandatory fields Association Type, Association ID and Association Source in the ASSOCIATION object uniquely identify the association group. If the optional TLVs - Global Association Source or Extended Association ID are included, then they MUST be included in combination with mandatory fields to uniquely identify the association group.

Association Information: As described in [RFC8697], the ASSOCIATION object could also include other TLVs based on the association types, that provides non-key information.

SRPAG: SR Policy Association Group.

SRPAT: SR Policy Association Type.

SRPAT ASSOCIATION: ASSOCIATION object of type SR Policy Association.

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

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

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

PCEP: Path Computation Element Protocol.

PCEP Tunnel: The entity identified by the PLSP-ID, as per [I-D.koldychev-pce-operational].

3. Motivation

The SR Policy Association and its TLVs, defined in this document, allow PCEP speakers to exchange additional information about SR Policy Candidate Paths and their container SR Policy.

3.1. Group Candidate Paths belonging to the same SR policy

Since each SR Policy Candidate Path appears as a different Tunnel (identified via a PLSP-ID) in PCEP, it is useful to group together all the SR Policy Candidate Paths that belong to the same SR Policy. Furthermore, it is useful for the PCE to have knowledge of the SR Policy related information such as color, endpoint, protocol origin, discriminator, and preference.

3.2. Instantiation of SR policy candidate paths

A PCE needs to instantiate one or more SR Policy Candidate Paths on the PCC, as specified in [RFC8281]. Each SR Policy Candidate Path is identified by the tuple <headend, color, endpoint, originator, discriminator, preference>. This draft provides a mechanism to signal this information in PCEP.

3.3. Avoid computing lower preference candidate paths

When a PCE knows that a given set of SR Policy Candidate Paths all belong to the same SR Policy, then path computation MAY be done on only the highest preference candidate-path(s). Path computation for lower preference paths is not necessary if one or two higher preference paths are already computed. Since computing their paths will not affect traffic steering, it MAY be postponed until the higher preference paths become invalid.

3.4. Minimal signaling overhead

When an SR Policy contains multiple SR Policy Candidate Paths computed by a PCE, such candidate paths can be created, updated and deleted independently of each other. This is achieved by making each SR Policy Candidate Path correspond to a unique Tunnel (identified via PLSP-ID). For example, if an SR Policy has 4 SR Policy Candidate Paths, then if the PCE wants to update one of those, only one set of PCUpd and PCRpt messages needs to be exchanged.

4. Procedure

4.1. Overview

As per [RFC8697], LSPs are placed into an association group. As per [I-D.koldychev-pce-operational], LSPs are contained in PCEP Tunnels and a PCEP Tunnel is contained in an Association if all of its LSPs are in that Association. PCEP Tunnels naturally map to SR Policy Candidate Paths and PCEP Associations naturally map to SR Policies.

The mapping between PCEP Associations and SR Policies is always one-to-one. However, the mapping between PCEP Tunnels and SR Policy Candidate Paths may be either one-to-one, or many-to-one, see Section 4.2.

Each SR Policy Candidate Path contains one or more Segment Lists. The subject of encoding multiple Segment Lists within an SR Policy Candidate Path is described in [I-D.koldychev-pce-multipath].

This document defines a new Association Type called "SR Policy Association", of value 6 based on the generic ASSOCIATION object. The new Association Type is also called "SRPAT", for "SR Policy Association Type". We say "SRPAT ASSOCIATION" to mean "ASSOCIATION object of type SR Policy Association". The group of LSPs that are part of the SR Policy Association is called "SRPAG", for "SR Policy Association Group".

As per the processing rules specified in section 6.4 of [RFC8697], if a PCEP speaker does not support the SRPAT, it MUST return a PCErr message with Error-Type = 26 "Association Error", Error-Value = 1 "Association-type is not supported".

A given LSP MUST belong to at most one SRPAG, since an SR Policy Candidate Path cannot belong to multiple SR Policies. If a PCEP speaker receives a PCEP message with more than one SRPAT ASSOCIATION for the same LSP, then the PCEP speaker MUST send a PCErr message with Error-Type = 26 "Association Error", Error-Value = 7 "Cannot join the association group".

An SRPAT ASSOCIATION carries three pieces of information: SR Policy Identifiers, SR Policy Candidate Path Identifiers, and SR Policy Candidate Path Attributes.

4.1.1. SR Policy Identifiers

SR Policy Identifiers uniquely identify the SR policy within the context of the headend. SR Policy Identifiers MUST be the same for all SR Policy Candidate Paths in the same SRPAG. SR Policy Identifiers MUST NOT change for a given SR Policy Candidate Path during its lifetime. SR Policy Identifiers MUST be different for different SRPAGs. SR Policy Identifiers consist of:

- * Headend router where the SR Policy originates.
- * Color of SR Policy.
- * Endpoint of SR Policy.

4.1.2. SR Policy Candidate Path Identifiers

SR Policy Candidate Path Identifiers uniquely identify the SR Policy Candidate Path within the context of an SR Policy. SR Policy Candidate Path Identifiers MUST NOT change for a given LSP during its lifetime. SR Policy Candidate Path Identifiers MUST be different for different LSPs within the same SRPAG. When these rules are not satisfied, the PCE MUST send a PCErr message with Error-Type = 26 "Association Error", Error Value = TBD8 "SR Policy Candidate Path Identifiers Mismatch". SR Policy Candidate Path Identifiers consist of:

- * Protocol Origin.
- * Originator.
- * Discriminator.

4.1.3. SR Policy Candidate Path Attributes

SR Policy Candidate Path Attributes carry non-key information about the candidate path and MAY change during the lifetime of the LSP. SR Policy Candidate Path Attributes consist of:

- * Preference.
- * Optionally, the SR Policy Candidate Path name.
- * Optionally, the SR Policy name.

4.2. Multiple Optimization Objectives and Constraints

In certain scenarios, it is desired for each SR Policy Candidate Path to contain multiple sub-candidate paths, each of which has a different optimization objective and constraints. Traffic is then sent ECMP or UCMP among these sub-candidate paths.

This is represented in PCEP by a many-to-one mapping between PCEP Tunnels and SR Policy Candidate Paths. This means that multiple PCEP Tunnels are allocated for each SR Policy Candidate Path. Each PCEP Tunnel has its own optimization objective and constraints. When a single SR Policy Candidate Path contains multiple PCEP Tunnels, each of these PCEP Tunnels MUST have identical values of Candidate Path Identifiers, as encoded in SRPOLICY-CPATH-ID TLV, see Section 5.2.2.

5. SR Policy Association

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 (6) "SR Policy Association". This Association Type is dynamic in nature, thus operator-configured Association Range MUST NOT be set for this Association type and MUST be ignored.

5.1. Association Parameters

As per [I-D.ietf-spring-segment-routing-policy], an SR Policy is identified through the tuple <headend, color, endpoint>. the headend is encoded as the Association Source in the ASSOCIATION object and the color and endpoint are encoded as part of Extended Association ID TLV.

The Association Parameters (see Section 2) consist of:

- * Association Type: set to 6 "SR Policy Association".
- * Association Source (IPv4/IPv6): set to the headend IP address.
- * Association ID (16-bit): set to "1".
- * Extended Association ID TLV: encodes the Color and Endpoint of the SR Policy.

The Association Source MUST be set to the headend value of the SR Policy, as defined in [I-D.ietf-spring-segment-routing-policy] Section 2.1. If the PCC receives a PCInit message for a non-existent SR Policy, where the Association Source is set not to the headend value but to some globally unique IP address that the PCC owns, then

the PCC SHOULD accept the PCInit message and create the SR Policy Association with the Association Source that was sent in the PCInit message.

The 16-bit Association ID field in the ASSOCIATION object MUST be set to the value of "1".

The Extended Association ID TLV MUST be included and it MUST be in the following format:

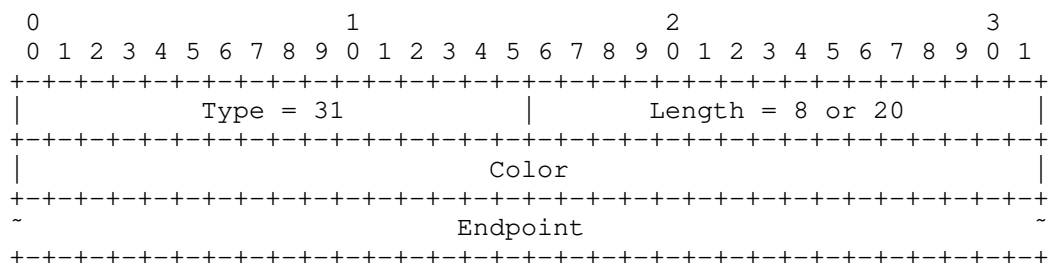


Figure 1: Extended Association ID TLV format

Type: Extended Association ID TLV, type = 31.

Length: Either 8 or 20, depending on whether IPv4 or IPv6 address is encoded in the Endpoint.

Color: SR Policy color value.

Endpoint: can be either IPv4 or IPv6, depending on whether the policy endpoint is IPv4 or IPv6. This value MAY be different from the one contained in the END-POINTS object, or in the LSP IDENTIFIERS TLV of the LSP object. This value is part of the tuple <color, endpoint> that identifies the SR Policy on a given headend.

If the PCEP speaker receives an SRPAT ASSOCIATION whose Association Parameters do not follow the above specification, then the PCEP speaker MUST send PCErr message with Error-Type = 26 "Association Error", Error-Value = TBD7 "SR Policy Identifiers Mismatch".

The purpose of choosing the Association Parameters in this way is to guarantee that there is no possibility of a race condition when multiple PCEP speakers want to create the same SR Policy at the same time. By adhering to this format, all PCEP speakers come up with the same Association Parameters independently of each other. Thus, there is no chance that different PCEP speakers will come up with different Association Parameters for the same SR Policy.

5.2. Association Information

The SRPAT ASSOCIATION contains the following TLVs:

- * SRPOLICY-POL-NAME TLV: (optional) encodes SR Policy Name string.
- * SRPOLICY-CPATH-ID TLV: (mandatory) encodes SR Policy Candidate Path Identifiers.
- * SRPOLICY-CPATH-NAME TLV: (optional) encodes SR Policy Candidate Path string name.
- * SRPOLICY-CPATH-PREFERENCE TLV: (optional) encodes SR Policy Candidate Path preference value.

Of these new TLVs, SRPOLICY-CPATH-ID TLV is mandatory. When a mandatory TLV is missing from the SRPAT ASSOCIATION object, the PCE MUST send a PCErr message with Error-Type = 6 "Mandatory Object Missing", Error-Value = TBD6 "Missing Mandatory TLV".

5.2.1. SR Policy Name TLV

The SRPOLICY-POL-NAME TLV is an optional TLV for the SRPAT ASSOCIATION. At most one SRPOLICY-POL-NAME TLV SHOULD be encoded by the sender and only the first occurrence is processed and any others MUST be ignored.

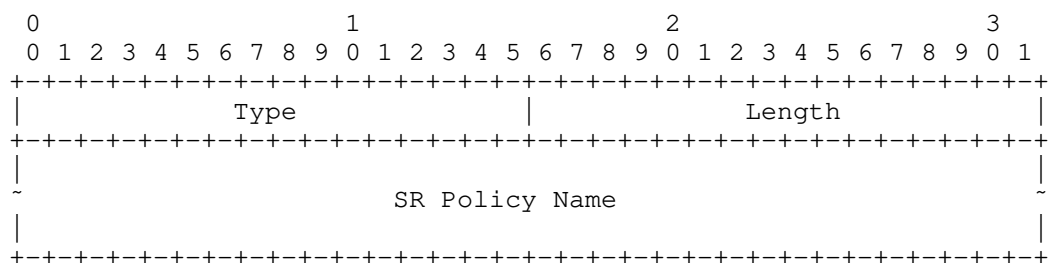


Figure 2: The SRPOLICY-POL-NAME TLV format

Type: 56 for "SRPOLICY-POL-NAME" TLV.

Length: indicates the length of the value portion of the TLV in octets and MUST be greater than 0. The TLV MUST be zero-padded so that the TLV is 4-octet aligned.

SR Policy Name: SR Policy name, as defined in [I-D.ietf-spring-segment-routing-policy]. It SHOULD be a string of printable ASCII characters, without a NULL terminator.

5.2.2. SR Policy Candidate Path Identifiers TLV

The SRPOLICY-CPATH-ID TLV is a mandatory TLV for the SRPAT ASSOCIATION. Only one SRPOLICY-CPATH-ID TLV SHOULD be encoded by the sender and only the first occurrence is processed and any others MUST be ignored.

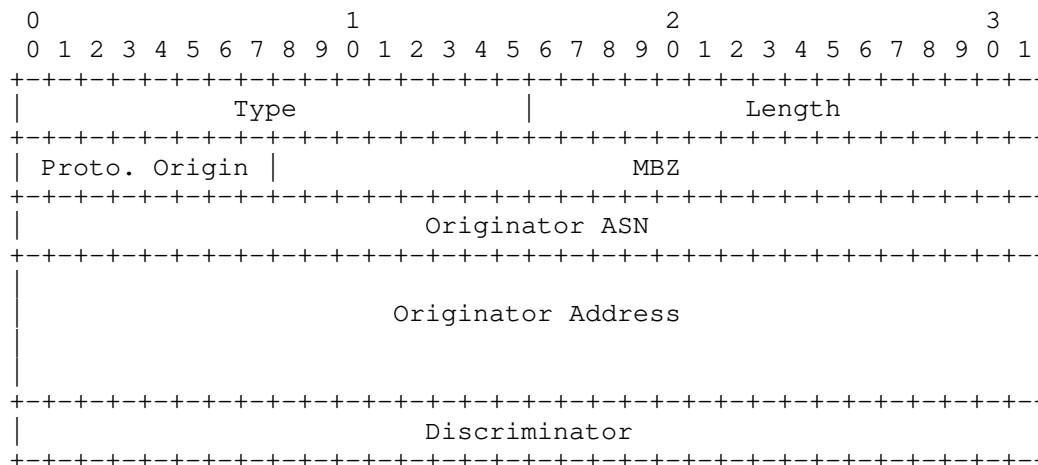


Figure 3: The SRPOLICY-CPATH-ID TLV format

Type: 57 for "SRPOLICY-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. Note that in PCInit messages, the Protocol Origin is always set to "PCEP".

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

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

5.2.3. SR Policy Candidate Path Name TLV

The SRPOLICY-CPATH-NAME TLV is an optional TLV for the SRPAT ASSOCIATION. At most one SRPOLICY-CPATH-NAME TLV SHOULD be encoded by the sender and only the first occurrence is processed and any others MUST be ignored.

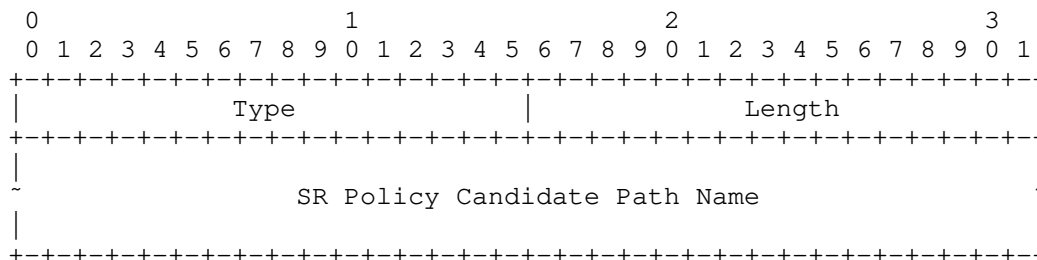


Figure 4: The SRPOLICY-CPATH-NAME TLV format

Type: 58 for "SRPOLICY-CPATH-NAME" TLV.

Length: indicates the length of the value portion of the TLV in octets and MUST be greater than 0. The TLV MUST be zero-padded so that the TLV is 4-octet aligned.

SR Policy Candidate Path Name: SR Policy Candidate Path Name, as defined in [I-D.ietf-spring-segment-routing-policy]. It SHOULD be a string of printable ASCII characters, without a NULL terminator.

5.2.4. SR Policy Candidate Path Preference TLV

The SRPOLICY-CPATH-PREFERENCE TLV is an optional TLV for the SRPAT ASSOCIATION. Only one SRPOLICY-CPATH-PREFERENCE TLV SHOULD be encoded by the sender and only the first occurrence is processed and any others MUST be ignored.

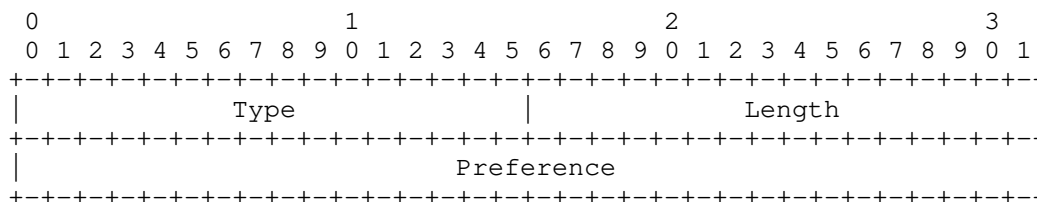


Figure 5: The SRPOLICY-CPATH-PREFERENCE TLV format

Type: 59 for "SRPOLICY-CPATH-PREFERENCE" TLV.

Length: 4.

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

If the TLV is missing, a default Preference value of 100 is used, as specified in Section 2.7 of [I-D.ietf-spring-segment-routing-policy].

6. Generic Mechanisms

This section describes various mechanisms that are standardized for SR Policies in [I-D.ietf-spring-segment-routing-policy], but are equally applicable to other tunnel types, such as RSVP-TE tunnels. Hence this section does not make use of the SRPAT ASSOCIATION.

6.1. Computation Priority TLV

The COMPUTATION-PRIORITY TLV is an optional TLV for the LSP object. It is used to signal the numerical computation priority, as specified in Section 2.12 of [I-D.ietf-spring-segment-routing-policy]. If the TLV is absent from the LSP object, a default Priority value of 128 is used.

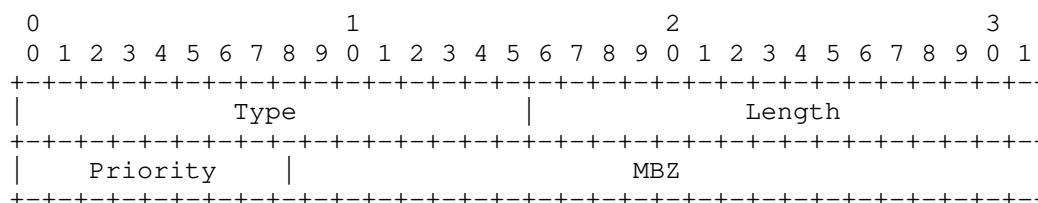


Figure 6: The COMPUTATION-PRIORITY TLV format

Type: TBD1 for "COMPUTATION-PRIORITY" TLV.

Length: 4.

Priority: Numerical priority with which this LSP is to be recomputed by the PCE upon topology change.

6.2. Explicit Null Label Policy (ENLP) TLV

The ENLP TLV is an optional TLV for the LSP object. It is used to implement the "Explicit Null Label Policy", as specified in Section 2.4.5 of [I-D.ietf-idr-segment-routing-te-policy].

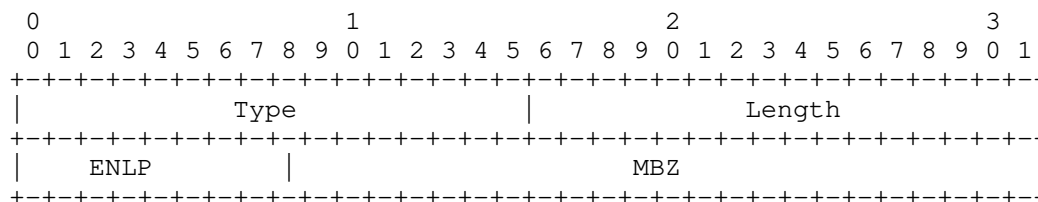


Figure 7: The Explicit Null Label Policy (ENLP) TLV format

Type: TBD2 for "ENLP" TLV.

Length: 4.

ENLP (Explicit NULL Label Policy): same values as in Section 2.4.5 of [I-D.ietf-idr-segment-routing-te-policy].

6.3. Invalidation TLV

The INVALIDATION TLV is an optional TLV for the LSP object. It is used to control traffic steering into the LSP during the time when the LSP is operationally down/invalid. In the context of SR Policy, this TLV facilitate the "Drop upon invalid" behavior, specified in Section 8.2 of [I-D.ietf-spring-segment-routing-policy]. Normally, if the LSP is down/invalid then traffic that is originally destined for that LSP is steered somewhere else, such as via IGP or via another LSP. The "Drop upon invalid" behavior specifies that such traffic MUST NOT be re-routed and has to be dropped at the head-end. While in the "Drop upon invalid" state, the LSP operational state is "UP", as indicated by the O-flag in the LSP object. However the ERO object is empty, indicating that traffic is being dropped.

In addition to the above, this TLV can also be used by the PCC to report to the PCE various reasons for LSP being invalidated. Invalidation reasons are represented by a set of flags.

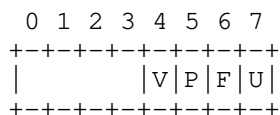


Figure 8: Invalidation Reasons Flags

- * U: Unknown - does not fit into any other categories below.
- * P: Path computation failure - no path was computed for the LSP.

- * F: First-hop resolution failure - head-end first hop resolution has failed.
- * V: Verification failure - OAM/PM/BFD path verification has indicated a breakage.

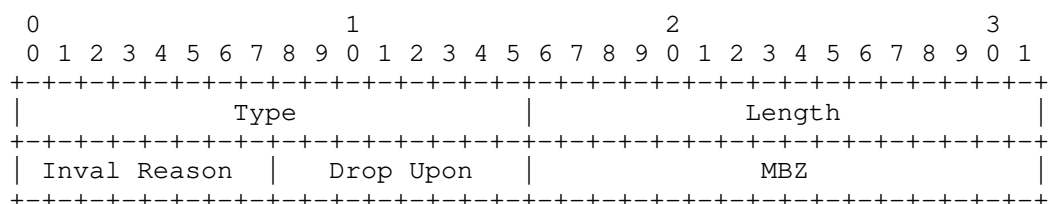


Figure 9: The INVALIDATION TLV format

Type: TBD3 for "INVALIDATION" TLV.

Length: 4.

Invalid Reason: contains "Invalidation Reasons Flags" which encode the reason(s) why the LSP is currently invalidated. This field can be set to non-zero values only by the PCC, it MUST be set to 0 by the PCE and ignored by the PCC.

Drop Upon: contains "Invalidation Reasons Flags" for conditions that SHOULD cause the LSP to drop traffic. This field can be set to non-zero values by both PCC and PCE. This field MAY be set to all 1's (0xFF) to indicate that the LSP is to go into Drop upon invalid state for any reason. I.e., when the PCE does not wish to distinguish any reason for LSP invalidation and just simply wants it to always "Drop upon invalid" for any reason.

6.4. Specified-BSID-only

Specified-BSID-only functionality is defined in Section 6.2.3 of [I-D.ietf-spring-segment-routing-policy]. When specified-BSID-only is enabled for a particular binding SID, it means that the given binding SID is required to be allocated and programmed for the LSP to be operationally up. If the binding SID cannot be allocated or programmed for some reason, then the LSP must stay down.

To signal specified-BSID-only, a new bit: S (Specified-BSID-only) is allocated in the "TE-PATH-BINDING TLV Flag field" of the TE-PATH-BINDING TLV. When this bit is set for a particular BSID, it means that the BSID follows the Specified-BSID-only behavior. It is possible to have a mix of BSIDs for the same LSP: some with S=1 and some with S=0.

7. Examples

7.1. PCC Initiated SR Policy with single candidate-path

PCReq and PCRep messages are exchanged in the following sequence:

1. PCC sends PCReq message to the PCE, encoding the SRPAT ASSOCIATION and TLVs in the PCReq message.
2. PCE returns the path in PCRep message, and echoes back the SRPAT ASSOCIATION.

PCRpt and PCUpd messages are exchanged in the following sequence:

1. PCC sends PCRpt message to the PCE, including the LSP object and the SRPAT ASSOCIATION.
2. PCE computes path, possibly making use of the Association Information from the SRPAT ASSOCIATION.
3. PCE updates the SR policy candidate path's ERO using PCUpd message.

7.2. PCC Initiated SR Policy with multiple candidate-paths

PCRpt and PCUpd messages are exchanged in the following sequence:

1. For each candidate path of the SR Policy, the PCC generates a different PLSP-ID and symbolic-name and sends multiple PCRpt messages (or one message with multiple LSP objects) to the PCE. Each LSP object is followed by SRPAT ASSOCIATION with identical Color and Endpoint values. The Association Source is set to the IP address of the PCC and the Association ID is set to a number that PCC locally chose to represent the SR Policy.
2. PCE takes into account that all the LSPs belong to the same SR policy. PCE prioritizes computation for the highest preference LSP and sends PCUpd message(s) back to the PCC.
3. If a new candidate path is added on the PCC by the operator, then a new PLSP-ID and symbolic name is generated for that candidate path and a new PCRpt is sent to the PCE.
4. If an existing candidate path is removed from the PCC by the operator, then that PLSP-ID is deleted from the PCE by sending PCRpt with the R-flag in the LSP object set.

7.3. PCE Initiated SR Policy with single candidate-path

A candidate-path is created using the following steps:

1. PCE sends PCInitiate message, containing the SRPAT ASSOCIATION. The Association Source and the Association ID are set as described in Section 5.1.
2. PCC uses the color, endpoint and preference from the SRPAT ASSOCIATION to create a new candidate path. If no SR policy exists to hold the candidate path, then a new SR policy is created to hold the new candidate-path. The Originator of the candidate path is set to be the address of the PCE that is sending the PCInitiate message.
3. PCC sends a PCRpt message back to the PCE to report the newly created Candidate Path. The PCRpt message contains the SRPAT ASSOCIATION.

A candidate-path is deleted using the following steps:

1. PCE sends PCInitiate message, setting the R-flag in the LSP object.
2. PCC uses the PLSP-ID from the LSP object to find the candidate path and delete it. If this is the last candidate path under the SR policy, then the containing SR policy is deleted as well.

7.4. PCE Initiated SR Policy with multiple candidate-paths

A candidate-path is created using the following steps:

1. PCE SHOULD send a separate PCInitiate message for every candidate path that it wants to create, or it MAY send multiple LSP objects within a single PCInitiate message. The SRPAT ASSOCIATION is sent for every LSP in the PCInitiate message. The Association Source and the Association ID are set as described in Section 5.1.
2. PCC creates multiple candidate paths under the same SR policy, identified by Color and Endpoint.
3. PCC sends a PCRpt message back to the PCE to report the newly created Candidate Path. The PCRpt message contains the SRPAT ASSOCIATION. The Association Source and the Association ID are set as described in Section 5.1.

A candidate path is deleted using the following steps:

1. PCE sends PCInitiate message, setting the R-flag in the LSP object.
2. PCC uses the PLSP-ID from the LSP object to find the candidate path and delete it.

8. IANA Considerations

8.1. Association Type

This document defines a new association type: SR Policy Association. IANA is requested to make the following codepoint assignment in the "ASSOCIATION Type Field" subregistry [RFC8697] within the "Path Computation Element Protocol (PCEP) Numbers" registry:

Type	Name	Reference
6	SR Policy Association	This.I-D

8.2. PCEP TLV Type Indicators

This document defines four new TLVs for carrying additional information about SR policy and SR candidate paths. IANA is requested to make the assignment of a new value for the existing "PCEP TLV Type Indicators" registry as follows:

Value	Description	Reference
56	SRPOLICY-POL-NAME	This.I-D
57	SRPOLICY-CPATH-ID	This.I-D
58	SRPOLICY-CPATH-NAME	This.I-D
59	SRPOLICY-CPATH-PREFERENCE	This.I-D
TBD1	COMPUTATION-PRIORITY	This.I-D
TBD2	EXPLICIT-NULL-LABEL-POLICY	This.I-D
TBD3	INVALIDATION	This.I-D

8.3. PCEP Errors

This document defines one new Error-Value within the "Mandatory Object Missing" Error-Type and two new Error-Values within the "Association Error" Error-Type. IANA is requested to allocate new error values within the "PCEP-ERROR Object Error Types and Values" sub-registry of the PCEP Numbers registry, as follows:

Error-Type	Meaning	Error-value	Reference
6	Mandatory Object Missing		[RFC5440]
		TBD6: SR Policy Missing Mandatory TLV	This.I-D
26	Association Error		[RFC8697]
		TBD7: SR Policy Identifiers Mismatch	This.I-D
		TBD8: SR Policy Candidate Path Identifiers Mismatch	This.I-D

8.4. TE-PATH-BINDING TLV Flag field

IANA is requested to allocate new bit within the "TE-PATH-BINDING TLV Flag field" sub-registry of the PCEP Numbers registry, as follows:

Bit position	Description	Reference
1	Specified-BSID-only	This.I-D

9. 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".

9.1. Cisco

- * Organization: Cisco Systems
- * Implementation: IOS-XR PCC and PCE.
- * Description: An experimental code-point is currently used.
- * Maturity Level: Proof of concept.
- * Coverage: Full.
- * Contact: mkoldych@cisco.com

9.2. Juniper

- * Organization: Juniper Networks
- * Implementation: Head-end and controller.
- * Description: An experimental code-point is currently used.
- * Maturity Level: Proof of concept.
- * Coverage: Partial.
- * Contact: cbarth@juniper.net

10. Security Considerations

This document defines one new type for association, which do not add any new security concerns beyond those discussed in [RFC5440], [RFC8231], [RFC8664], [I-D.ietf-pce-segment-routing-ipv6] and [RFC8697] in itself.

The information carried in the SRPAT ASSOCIATION, as per this document is related to SR Policy. It often reflects information that can also be derived from the SR Database, but association provides a much easier grouping of related LSPs and messages. The SRPAT ASSOCIATION could provide an adversary with the opportunity to eavesdrop on the relationship between the LSPs. Thus securing the PCEP session using Transport Layer Security (TLS) [RFC8253], as per the recommendations and best current practices in [RFC7525], is RECOMMENDED.

11. Acknowledgement

Would like to thank Stephane Litkowski, Boris Khasanov, Praveen Kumar and Tom Petch for review and suggestions.

12. References

12.1. Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <<https://www.rfc-editor.org/info/rfc2119>>.
- [RFC5440] Vasseur, JP., Ed. and JL. Le Roux, Ed., "Path Computation Element (PCE) Communication Protocol (PCEP)", RFC 5440, DOI 10.17487/RFC5440, March 2009, <<https://www.rfc-editor.org/info/rfc5440>>.
- [RFC8174] Leiba, B., "Ambiguity of Uppercase vs Lowercase in RFC 2119 Key Words", BCP 14, RFC 8174, DOI 10.17487/RFC8174, May 2017, <<https://www.rfc-editor.org/info/rfc8174>>.
- [RFC8231] Crabbe, E., Minei, I., Medved, J., and R. Varga, "Path Computation Element Communication Protocol (PCEP) Extensions for Stateful PCE", RFC 8231, DOI 10.17487/RFC8231, September 2017, <<https://www.rfc-editor.org/info/rfc8231>>.

- [RFC8281] Crabbe, E., Minei, I., Sivabalan, S., and R. Varga, "Path Computation Element Communication Protocol (PCEP) Extensions for PCE-Initiated LSP Setup in a Stateful PCE Model", RFC 8281, DOI 10.17487/RFC8281, December 2017, <<https://www.rfc-editor.org/info/rfc8281>>.
- [RFC7942] Sheffer, Y. and A. Farrel, "Improving Awareness of Running Code: The Implementation Status Section", BCP 205, RFC 7942, DOI 10.17487/RFC7942, July 2016, <<https://www.rfc-editor.org/info/rfc7942>>.
- [I-D.ietf-spring-segment-routing-policy] Filsfils, C., Talaulikar, K., Voyer, D., Bogdanov, A., and P. Mattes, "Segment Routing Policy Architecture", Work in Progress, Internet-Draft, draft-ietf-spring-segment-routing-policy-22, 22 March 2022, <<https://www.ietf.org/archive/id/draft-ietf-spring-segment-routing-policy-22.txt>>.
- [I-D.ietf-idr-segment-routing-te-policy] Previdi, S., Filsfils, C., Talaulikar, K., Mattes, P., Jain, D., and S. Lin, "Advertising Segment Routing Policies in BGP", Work in Progress, Internet-Draft, draft-ietf-idr-segment-routing-te-policy-17, 14 April 2022, <<https://www.ietf.org/archive/id/draft-ietf-idr-segment-routing-te-policy-17.txt>>.
- [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>>.
- [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>>.
- [I-D.koldychev-pce-operational] Koldychev, M., Sivabalan, S., Peng, S., Achaval, D., and H. Kotni, "PCEP Operational Clarification", Work in Progress, Internet-Draft, draft-koldychev-pce-operational-05, 19 February 2022, <<https://www.ietf.org/archive/id/draft-koldychev-pce-operational-05.txt>>.

[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", Work in Progress, Internet-Draft, draft-koldychev-pce-multipath-05, 16 February 2021, <<https://www.ietf.org/archive/id/draft-koldychev-pce-multipath-05.txt>>.

12.2. Informative References

- [RFC4655] Farrel, A., Vasseur, J.-P., and J. Ash, "A Path Computation Element (PCE)-Based Architecture", RFC 4655, DOI 10.17487/RFC4655, August 2006, <<https://www.rfc-editor.org/info/rfc4655>>.
- [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>>.
- [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>>.
- [I-D.ietf-pce-segment-routing-ipv6]
Li, C., Negi, M., Sivabalan, S., Koldychev, M., Kaladharan, P., and Y. Zhu, "PCEP Extensions for Segment Routing leveraging the IPv6 data plane", Work in Progress, Internet-Draft, draft-ietf-pce-segment-routing-ipv6-13, 1 April 2022, <<https://www.ietf.org/internet-drafts/draft-ietf-pce-segment-routing-ipv6-13.txt>>.

Appendix A. Contributors

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

Email: dhruv.ietf@gmail.com

Cheng Li
Huawei Technologies
Huawei Campus, No. 156 Beiqing Rd.
Beijing, 10095
China

Email: chengli13@huawei.com

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

Email: ssidor@cisco.com

Authors' Addresses

Mike Koldychev
Cisco Systems, Inc.
2000 Innovation Drive
Kanata Ontario K2K 3E8
Canada
Email: mkoldych@cisco.com

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

Colby Barth
Juniper Networks, Inc.
Email: cbarth@juniper.net

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

Hooman Bidgoli
Nokia
Email: hooman.bidgoli@nokia.com

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.

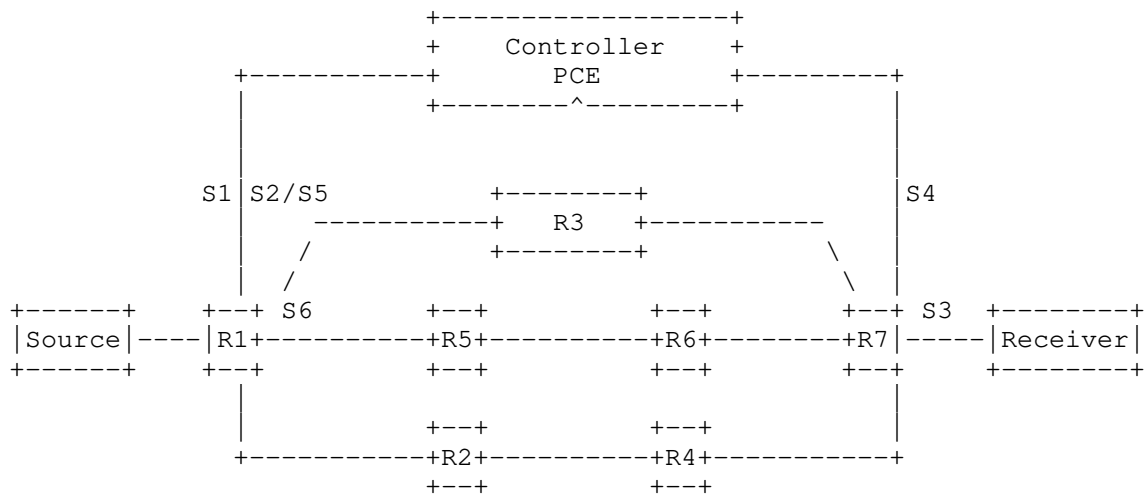


Figure 1: Example of PCE based BIER Topology(controller as PCE)

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:

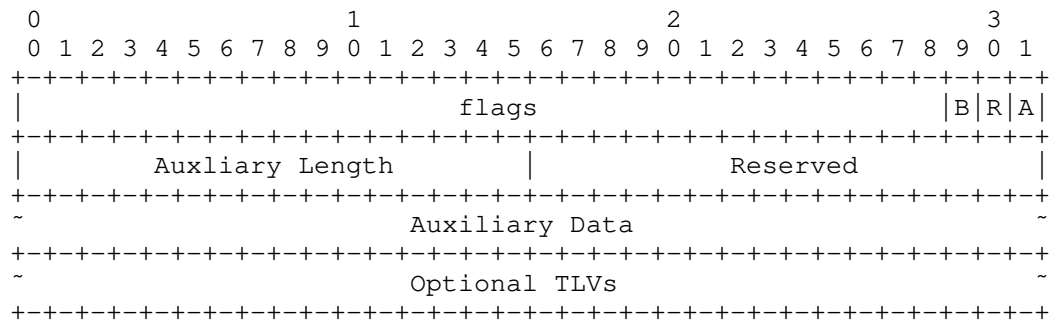


Figure 2: MSR Object Body Format

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:

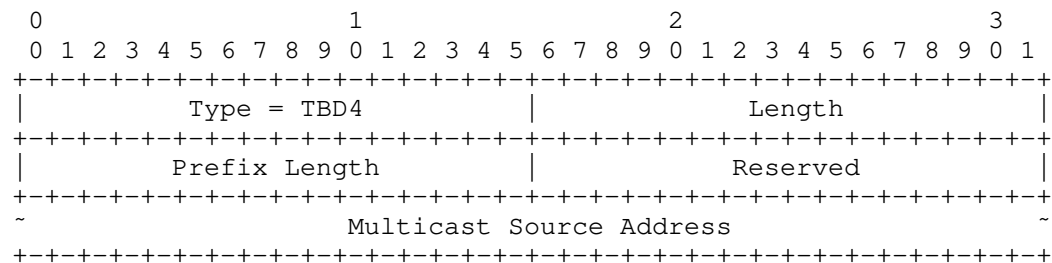


Figure 3: Multicast Source Address TLV Format

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:

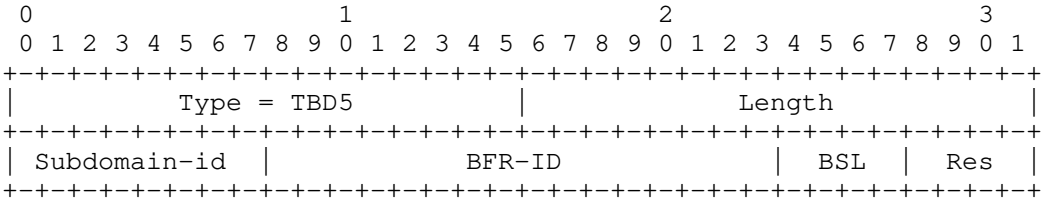


Figure 4: BIER Information TLV Format

Type(16 bits): TBD5 is to be assigned by IANA.

Length: Variable.

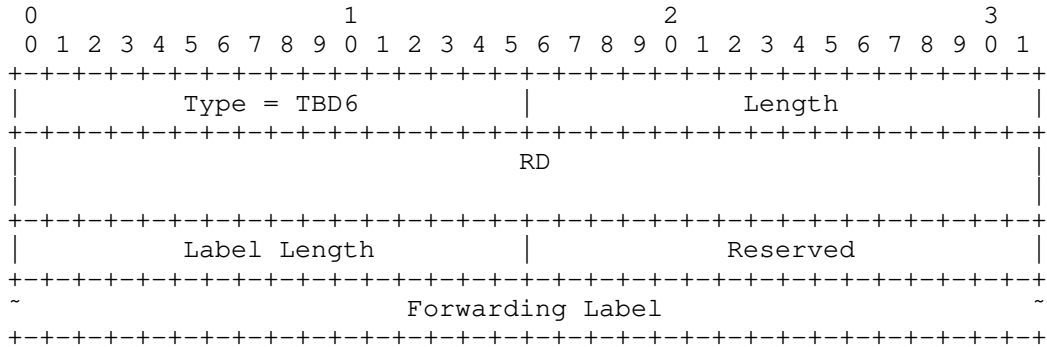
Subdomain-id(8 bits): Unique value identifying the BIER subdomain.

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:



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.

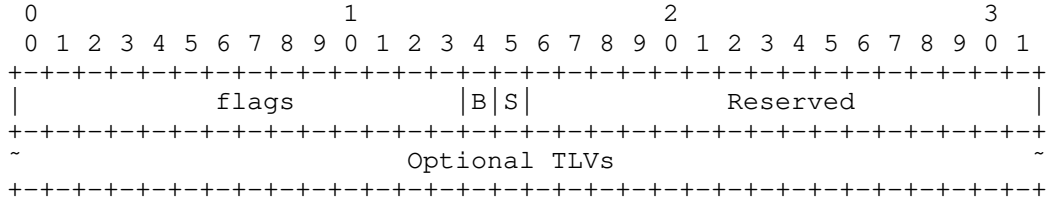
Forwarding Label(Variable Length): contains MPLS label with 32 bit or IPv6 Segment Identifier with 128 bits.

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 muticast joining or leaving.

MRI Object-Class is TBD7. MRI Object-Type is 1.

The format of the MRI object body is:



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:

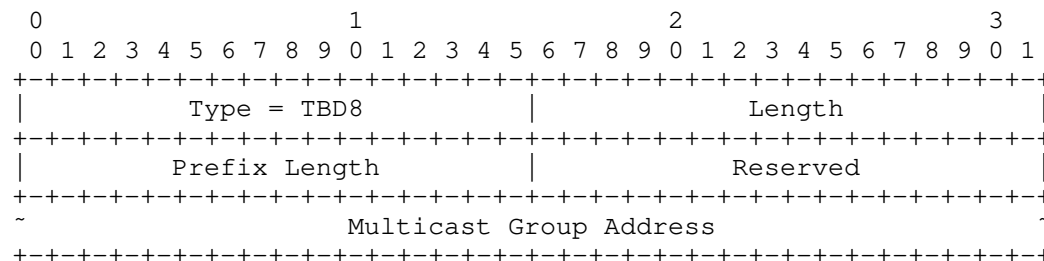


Figure 7: Multicast Group Address TLV Format

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:

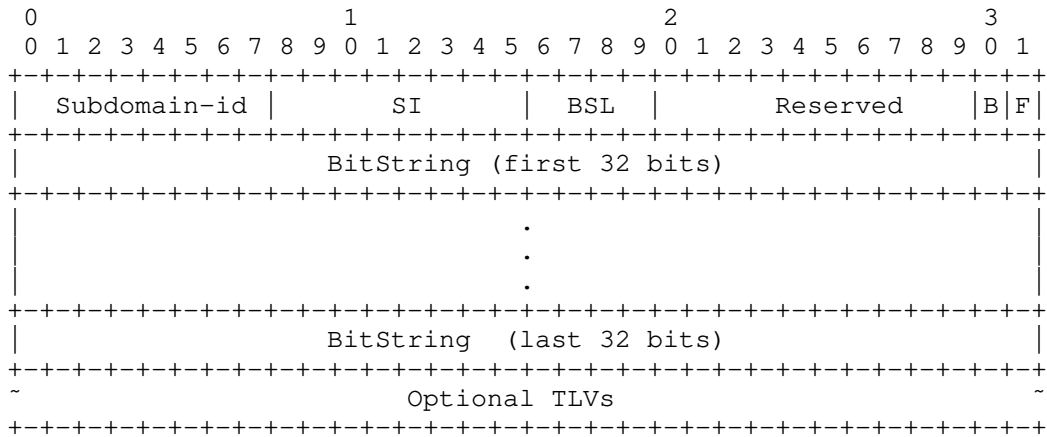


Figure 8: FI Object Body Format

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:

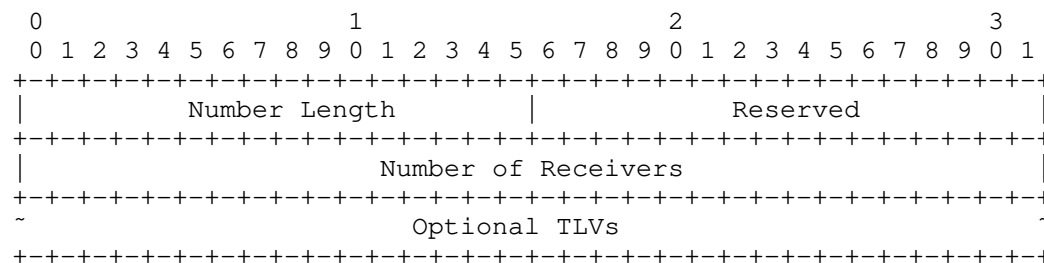


Figure 9: MRS Object Body Format

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 TLVSection 7.1.1 and Multicast Group Address TLVSection 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 Working Group
Internet-Draft
Intended status: Standards Track
Expires: 24 April 2022

S. Peng
C. Li
Huawei Technologies
L. Han
China Mobile
L. Ndifor
MTN Cameroon
21 October 2021

Support for Path MTU (PMTU) in the Path Computation Element (PCE)
communication Protocol (PCEP).
draft-li-pce-pcep-pmtu-05

Abstract

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

The Source Packet Routing in Networking (SPRING) architecture describes how Segment Routing (SR) can be used to steer packets through an IPv6 or MPLS network using the source routing paradigm. A Segment Routed Path can be derived from a variety of mechanisms, including an IGP Shortest Path Tree (SPT), explicit configuration, or a Path Computation Element (PCE).

Since the SR does not require signaling, the path maximum transmission unit (MTU) information for SR path is not available. This document specifies the extension to PCE communication protocol (PCEP) to carry path (MTU) in the PCEP messages.

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 24 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	4
3. PCEP Extension	5
3.1. Extensions to METRIC Object	5
3.2. Multi-Path Handling	6
3.3. Stateful PCE and PCE Initiated LSPs	7
3.4. Segment Routing	7
3.5. Path MTU Adjustment	7
4. Security Considerations	8
5. IANA Considerations	8
5.1. METRIC Type	8
6. Acknowledgement	8
7. References	8
7.1. Normative References	8
7.2. Informative References	9
Authors' Addresses	11

1. Introduction

[RFC5440] describes the Path Computation Element (PCE) Communication Protocol (PCEP). PCEP enables the communication between a Path Computation Client (PCC) and a PCE, or between PCE and PCE, for the purpose of computation of Multiprotocol Label Switching (MPLS) as well as Generalized MPLS (GMPLS) Traffic Engineering Label Switched Path (TE LSP) characteristics.

[RFC8231] specifies a set of extensions to PCEP to enable stateful control of TE LSPs within and across PCEP sessions in compliance with [RFC4657]. It includes mechanisms to effect LSP State Synchronization between PCCs and PCEs, delegation of control over LSPs to PCEs, and PCE control of timing and sequence of path computations within and across PCEP sessions. The model of operation where LSPs are initiated from the PCE is described in [RFC8281].

As per [RFC8402], with Segment Routing (SR), a node steers a packet through an ordered list of instructions, called segments. A segment can represent any instruction, topological or service-based. A segment can have a semantic local to an SR node or global within an SR domain. SR allows to enforce a flow through any path and service chain while maintaining per-flow state only at the ingress node of the SR domain. Segments can be derived from different components: IGP, BGP, Services, Contexts, Locators, etc. The SR architecture can be applied to the MPLS forwarding plane without any change, in which case an SR path corresponds to an MPLS Label Switching Path (LSP). The SR is applied to IPv6 forwarding plane using SRH. A SR path can be derived from an IGP Shortest Path Tree (SPT), but SR-TE paths may not follow IGP SPT. Such paths may be chosen by a suitable network planning tool, or a PCE and provisioned on the ingress node.

As per [RFC8664], it is possible to use a stateful PCE for computing one or more SR-TE paths taking into account various constraints and objective functions. Once a path is chosen, the stateful PCE can initiate an SR-TE path on a PCC using PCEP extensions specified in [RFC8281] using the SR specific PCEP extensions specified in [RFC8664]. [RFC8664] specifies PCEP extensions for supporting a SR-TE LSP for MPLS data plane. [I-D.ietf-pce-segment-routing-ipv6] extend PCEP to support SR for IPv6 data plane.

The maximum transmission unit (MTU) is the largest size packet or frame, in bytes, that can be sent in a network. An MTU that is too large might cause retransmissions. Too small an MTU might cause the router to send and handle relatively more header overhead and acknowledgments. When an LSP is created across a set of links with different MTU sizes, the ingress router needs to know what the smallest MTU is on the LSP path. If this MTU is larger than the MTU

of one of the intermediate links, traffic might be dropped, because MPLS packets cannot be fragmented. Also, the ingress router may not be aware of this type of traffic loss, because the control plane for the LSP would still function normally. [RFC3209] specify the mechanism of MTU signaling in RSVP.

Since the SR does not require signaling, the path MTU information for SR path is not available. This document specify the extension to PCEP to carry path MTU in the PCEP messages. It is assumed that the PCE is aware of the link MTU as part of the Traffic Engineering Database (TED) population. This could be done via IGP, BGP-LS [I-D.ietf-idr-bgp-ls-link-mtu] or some other means. Thus the PCE can find the path MTU at the time of path computation and include this information as part of the PCEP messages.

Though the key use case for path MTU is SR, the PCEP extension (as specified in this document) creates a new metric type for path MTU, making this a generic extension that can be used independent of SR.

Note that in SR, the term Maximum SID Depth (MSD) [RFC8491] refers to the maximum number of SIDs that an ingress is capable of imposing on a packet. The PMTU on the other hand determines if the IP fragmentation could be avoided.

2. Terminology

This draft refers to the terms defined in [RFC8201], [RFC4821] and [RFC3988].

MTU: Maximum Transmission Unit, the size in bytes of the largest IP packet, including the IP header and payload, that can be transmitted on a link or path. Note that this could more properly be called the IP MTU, to be consistent with how other standards organizations use the acronym MTU.

Link MTU: The Maximum Transmission Unit, i.e., maximum IP packet size in bytes, that can be conveyed in one piece over a link. Be aware that this definition is different from the definition used by other standards organizations.

For IETF documents, link MTU is uniformly defined as the IP MTU over the link. This includes the IP header, but excludes link layer headers and other framing that is not part of IP or the IP payload.

Be aware that other standards organizations generally define link MTU to include the link layer headers.

For the MPLS data plane, this size includes the IP header and data (or other payload) and the label stack but does not include any lower-layer headers. A link may be an interface (such as Ethernet or Packet-over-SONET), a tunnel (such as GRE or IPsec), or an LSP.

Path: The set of links traversed by a packet between a source node and a destination node.

Path MTU, or PMTU: The minimum link MTU of all the links in a path between a source node and a destination node.

For the MPLS data plane, it is the MTU of an LSP from a given LSR to the egress(es), over each valid (forwarding) path. This size includes the IP header and data (or other payload) and any part of the label stack that was received by the ingress LSR before it placed the packet into the LSP (this part of the label stack is considered part of the payload for this LSP). The size does not include any lower-level headers.

3. PCEP Extention

3.1. Extensions to METRIC Object

The METRIC object is defined in Section 7.8 of [RFC5440], comprising metric-value and metric-type (T field), and a flags field, comprising a number of bit flags (B bit and C bit). This document defines a new type for the METRIC object for Path MTU.

* T = TBD: Path MTU.

- * A network comprises of a set of N links $\{L_i, (i=1...N)\}$.
- * A path P of a LSP is a list of K links $\{L_{pi}, (i=1...K)\}$.
- * A Link MTU of link L is denoted $M(L)$.
- * A Path MTU metric for the path $P = \text{Min } \{M(L_{pi}), (i=1...K)\}$.

The Path MTU metric type of the METRIC object in PCEP represents the minimum of the Link MTU of all links along the path.

When PCE computes the path, it can also find the Path MTU (based on the above criteria) and include this information in the METRIC object with the above metric type in the PCEP message when replying to the PCC. In a Path Computation Reply (PCRep) message, the PCE MAY insert the METRIC object with an Explicit Route Object (ERO) so as to provide the METRIC (path MTU) for the computed path. The PCE MAY also insert the METRIC object with a NO-PATH object to indicate that the metric constraint could not be satisfied.

Further, a PCC MAY use the Path MTU metric in a Path Computation Request (PCReq) message to request a path meeting the MTU requirement of the path. In this case, the B bit MUST be set to suggest a bound (a maximum) for the Path MTU metric that must not be exceeded for the PCC to consider the computed path as acceptable. The Path MTU metric must be less than or equal to the value specified in the metric-value field.

A PCC can also use this metric to ask PCE to optimize the path MTU during path computation. In this case, the B bit MUST be cleared.

The error handling and processing of the METRIC object is as specified in [RFC5440].

3.2. Multi-Path Handling

[I-D.ietf-pce-multipath] extends PCEP to support signaling of multipath information i.e. to all each Candidate-Path to contain multiple Segment-Lists.

The PMTU could be supported per segment list as well. The exact mechanism to support this is left for further revision of this document.

3.3. Stateful PCE and PCE Initiated LSPs

[RFC8231] specifies a set of extensions to PCEP to enable stateful control of MPLS-TE LSPs via PCEP and the maintaining of these LSPs at the stateful PCE. It further distinguishes between an active and a passive stateful PCE. A passive stateful PCE uses LSP state information learned from PCCs to optimize path computations but does not actively update LSP state. In contrast, an active stateful PCE utilizes the LSP delegation mechanism to update LSP parameters in those PCCs that delegated control over their LSPs to the PCE. [RFC8281] describes the setup, maintenance, and teardown of PCE-initiated LSPs under the stateful PCE model. The document defines the PCInitiate message that is used by a PCE to request a PCC to set up a new LSP.

The new metric type defined in this document can also be used with the stateful PCE extensions. The format of PCEP messages described in [RFC8231] and [RFC8281] uses <intended-attribute-list> and <attribute-list>, respectively, (where the <intended-attribute-list> is the attribute-list defined in Section 6.5 of [RFC5440]).

A PCE MAY include the path MTU metric in PCInitiate or PCUpd message to inform the PCC of the path MTU calculated for the path. A PCC MAY include the path MTU metric as a bound constraint or to indicate optimization criteria (similar to PCReq).

3.4. Segment Routing

A Segment Routed path (SR path) can be derived from an IGP Shortest Path Tree (SPT). Segment Routed Traffic Engineering paths (SR-TE paths) may not follow IGP SPT. Such paths may be chosen by a suitable network planning tool and provisioned on the source node of the SR-TE path.

It is possible to use a PCE for computing one or more SR-TE paths taking into account various constraints and objective functions. Once a path is chosen, the PCE can inform an SR-TE path on a PCC using PCEP extensions specified in [RFC8664]. Further, [I-D.ietf-pce-segment-routing-ipv6] adds the support for IPv6 data plane in SR.

The new metric type for path MTU is applicable for the SR-TE path and require no additional extensions.

3.5. Path MTU Adjustment

The path MTU metric can be used for both primary and protection path.

The minimal value of the link MTU along the path is collected, based on which minor adjustment is made to cater for overhead introduced by the protection mechanisms such as TI-LFA. The path MTU is the value of the minimum link MTU minus the overhead. In this way, the ingress node can use the path MTU directly.

4. Security Considerations

This document defines a new METRIC type that do not add any new security concerns beyond those discussed in [RFC5440] in itself. Some deployments may find the path MTU information to be extra sensitive and could be used to influence path computation and setup with adverse effect. Additionally, snooping of PCEP messages with such data or using PCEP messages for network reconnaissance may give an attacker sensitive information about the operations of the network. Thus, such deployment should employ suitable PCEP security mechanisms like TCP Authentication Option (TCP-AO) [RFC5925] or Transport Layer Security (TLS) [RFC8253]. The procedure based on TLS is considered a security enhancement and thus is much better suited for the sensitive information.

5. IANA Considerations

This document makes following requests to IANA for action.

5.1. METRIC Type

IANA maintains the "Path Computation Element Protocol (PCEP) Numbers" registry. Within this registry, IANA maintains a subregistry for "METRIC Object T Field". IANA is requested to make the following allocation:

Value	Description	Reference
TBD	Path MTU	This document

6. Acknowledgement

We would like to thank Dhruv Dhody for his contributions for this document.

7. References

7.1. Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <<https://www.rfc-editor.org/info/rfc2119>>.
- [RFC5440] Vasseur, JP., Ed. and JL. Le Roux, Ed., "Path Computation Element (PCE) Communication Protocol (PCEP)", RFC 5440, DOI 10.17487/RFC5440, March 2009, <<https://www.rfc-editor.org/info/rfc5440>>.
- [RFC8174] Leiba, B., "Ambiguity of Uppercase vs Lowercase in RFC 2119 Key Words", BCP 14, RFC 8174, DOI 10.17487/RFC8174, May 2017, <<https://www.rfc-editor.org/info/rfc8174>>.
- [RFC8231] Crabbe, E., Minei, I., Medved, J., and R. Varga, "Path Computation Element Communication Protocol (PCEP) Extensions for Stateful PCE", RFC 8231, DOI 10.17487/RFC8231, September 2017, <<https://www.rfc-editor.org/info/rfc8231>>.
- [RFC8281] Crabbe, E., Minei, I., Sivabalan, S., and R. Varga, "Path Computation Element Communication Protocol (PCEP) Extensions for PCE-Initiated LSP Setup in a Stateful PCE Model", RFC 8281, DOI 10.17487/RFC8281, December 2017, <<https://www.rfc-editor.org/info/rfc8281>>.

7.2. Informative References

- [RFC3209] Awduche, D., Berger, L., Gan, D., Li, T., Srinivasan, V., and G. Swallow, "RSVP-TE: Extensions to RSVP for LSP Tunnels", RFC 3209, DOI 10.17487/RFC3209, December 2001, <<https://www.rfc-editor.org/info/rfc3209>>.
- [RFC3988] Black, B. and K. Kompella, "Maximum Transmission Unit Signalling Extensions for the Label Distribution Protocol", RFC 3988, DOI 10.17487/RFC3988, January 2005, <<https://www.rfc-editor.org/info/rfc3988>>.
- [RFC4657] Ash, J., Ed. and J.L. 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>>.
- [RFC4821] Mathis, M. and J. Heffner, "Packetization Layer Path MTU Discovery", RFC 4821, DOI 10.17487/RFC4821, March 2007, <<https://www.rfc-editor.org/info/rfc4821>>.

- [RFC5925] Touch, J., Mankin, A., and R. Bonica, "The TCP Authentication Option", RFC 5925, DOI 10.17487/RFC5925, June 2010, <<https://www.rfc-editor.org/info/rfc5925>>.
- [RFC8201] McCann, J., Deering, S., Mogul, J., and R. Hinden, Ed., "Path MTU Discovery for IP version 6", STD 87, RFC 8201, DOI 10.17487/RFC8201, July 2017, <<https://www.rfc-editor.org/info/rfc8201>>.
- [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>>.
- [RFC8402] Filts, 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>>.
- [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>>.
- [RFC8664] Sivabalan, S., Filts, 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>>.
- [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", Work in Progress, Internet-Draft, draft-ietf-pce-multipath-02, 17 October 2021, <<https://www.ietf.org/archive/id/draft-ietf-pce-multipath-02.txt>>.
- [I-D.ietf-pce-segment-routing-ipv6]
Li, C., Negi, M., Sivabalan, S., Koldychev, M., Kaladharan, P., and Y. Zhu, "PCEP Extensions for Segment Routing leveraging the IPv6 data plane", Work in Progress, Internet-Draft, draft-ietf-pce-segment-routing-ipv6-09, 27 May 2021, <<https://www.ietf.org/internet-drafts/draft-ietf-pce-segment-routing-ipv6-09.txt>>.

[I-D.ietf-idr-bgp-ls-link-mtu]

Zhu, Y., Hu, Z., Peng, S., and R. Mwehaire, "Signaling Maximum Transmission Unit (MTU) using BGP-LS", Work in Progress, Internet-Draft, draft-ietf-idr-bgp-ls-link-mtu-01, 25 May 2021, <<https://www.ietf.org/archive/id/draft-ietf-idr-bgp-ls-link-mtu-01.txt>>.

Authors' Addresses

Shuping Peng
Huawei Technologies
Huawei Campus, No. 156 Beiqing Rd.
Beijing
100095
China

Email: pengshuping@huawei.com

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

Email: c.l@huawei.com

Liuyan Han
China Mobile
Beijing
100053
China

Email: hanliuyan@chinamobile.com

Luc-Fabrice Ndifor
MTN Cameroon
Cameroon

Email: Luc-Fabrice.Ndifor@mtn.com

PCE Working Group
Internet-Draft
Intended status: Standards Track
Expires: 28 April 2022

B. Rajagopalan
V. Beeram
Juniper Networks
S. Peng
Q. Xiong
ZTE Corporation
M. Koldychev
Cisco Systems Inc.
G. Mishra
Verizon Communications Inc.
25 October 2021

Path Computation Element Protocol(PCEP) Extension for Color
draft-rajagopalan-pce-pcep-color-00

Abstract

Color is a 32-bit numerical attribute that is used to associate a Traffic Engineering (TE) tunnel or policy with an intent or objective (e.g. low latency). This document specifies an extension to Path Computation Element Protocol (PCEP) to carry the color attribute.

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 28 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	2
2. Use case: RSVP-TE Color	3
3. Protocol Operation	3
4. TLV Format	4
5. Security Considerations	5
6. IANA Considerations	5
6.1. PCEP TLV Type Indicator	5
6.2. STATEFUL-PCE-CAPABILITY TLV Flag Field	5
6.3. LSP-ERROR-CODE TLV Error Code Field	5
7. Acknowledgments	6
8. References	6
8.1. Normative References	6
8.2. Informative References	7
Authors' Addresses	8

1. Introduction

A Traffic Engineering (TE) tunnel or policy can be associated with an intent or objective (e.g. low latency) by marking it with a color. This color attribute is used as a guiding criterion for mapping services onto the TE tunnel or policy ([RFC9012]). The term color used in this document is NOT to be interpreted as the 'thread color' specified in [RFC3063] or the 'resource color' (or 'link color') specified in [RFC3630], [RFC5329], [RFC5305] and [RFC7308].

Color is part of the tuple that identifies a Segment Routing (SR) policy ([I-D.ietf-spring-segment-routing-policy]) and is included in the Path Computation Element Protocol (PCEP) extensions defined for carrying the SR policy identifiers ([I-D.ietf-pce-segment-routing-policy-cp]). The color encoding specified in SR policy identifier cannot be reused for other types of path setup.

This document introduces a generic optional PCEP TLV called the Color TLV to carry the color attribute and discusses its usage with RSVP-TE Label Switched Paths (LSPs).

In addition to catering to the use-case discussed in this document, the Color TLV can also be used to reference SR Composite Candidate Paths as specified in ([I-D.ietf-pce-multipath]). An implementation MAY also provide a local policy option to use this TLV to reference a set of path constraints and optimization objectives.

2. Use case: RSVP-TE Color

The color attribute can be used as one of the guiding criteria in selecting the RSVP-TE LSP as a next hop for service prefixes. While the specific details of how the service prefixes are associated with the appropriate RSVP-TE LSPs are outside the scope of this specification, the envisioned high level usage of the color attribute is as follows.

The service prefixes are marked with some indication of the type of underlay they need. The underlay LSPs 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 LSPs.

As an example, for a BGP-based service, the originating PE could attach some community, e.g. the Extended Color Community [RFC9012] with the service route. A receiving PE could use locally configured policies to associate service routes carrying Extended Color Community 'X' with underlay RSVP-TE LSPs 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-TE LSPs of a certain color.

The procedure discussed for service mapping in this section can be applied to any underlay path setup type.

3. Protocol Operation

The STATEFUL-PCE-CAPABILITY negotiation message is enhanced to carry the color capability, which allows PCC (Path Computation Client) and PCE (Path Computation Element) 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

- * SHOULD NOT report color to the PCE.
- * 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" (TBA3) in LSP State Report message.

When LSPs that belong to the same TE tunnel are with in the same Path Protection Association Group [RFC8745], the color is attached only to the primary LSP. If PCC receives color TLV for a secondary LSP, it SHOULD respond with an error code of 4 (Unacceptable Parameters).

4. TLV Format

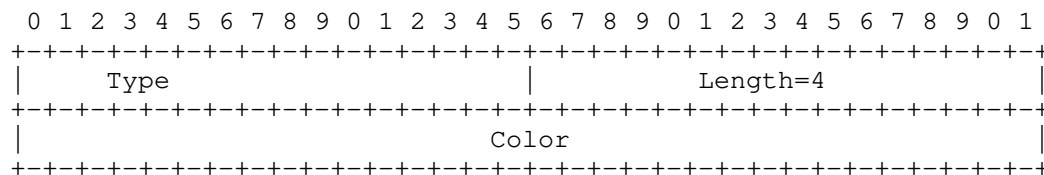


Figure 1: Color TLV

Type has the value TBA1. 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 (TBA2): A PCE/PCC that supports color capability must turn on this bit.

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

6.1. PCEP TLV Type Indicator

IANA is requested to allocate a new value in the "PCEP TLV Type Indicators" sub-registry of the PCEP Numbers registry as follows:

Value	Description	Reference
TBA1	Color	This document

6.2. STATEFUL-PCE-CAPABILITY TLV Flag Field

IANA is requested to allocate a new bit value in the "STATEFUL-PCE-CAPABILITY TLV Flag Field" sub-registry of the PCEP Numbers registry as follows:

Value	Description	Reference
TBA2	COLOR-CAPABILITY	This document

6.3. LSP-ERROR-CODE TLV Error Code Field

IANA is requested to allocate a new error code in the "LSP-ERROR-CODE TLV Error Code Field" sub-registry of the PCEP Numbers registry as follows:

Value	Meaning	Reference
TBA3	Unsupported Color	This document

7. Acknowledgments

The authors would like to thank Kaliraj Vairavakkalai, Colby Barth, Natrajan Venkataraman and Tarek Saad for their review and suggestions.

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>>.
- [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>>.

- [RFC8745] Ananthakrishnan, H., Sivabalan, S., Barth, C., Minei, I., and M. Negi, "Path Computation Element Communication Protocol (PCEP) Extensions for Associating Working and Protection Label Switched Paths (LSPs) with Stateful PCE", RFC 8745, DOI 10.17487/RFC8745, March 2020, <<https://www.rfc-editor.org/info/rfc8745>>.
- [RFC9012] Patel, K., Van de Velde, G., Sangli, S., and J. Scudder, "The BGP Tunnel Encapsulation Attribute", RFC 9012, DOI 10.17487/RFC9012, April 2021, <<https://www.rfc-editor.org/info/rfc9012>>.

8.2. Informative References

- [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", Work in Progress, Internet-Draft, draft-ietf-pce-multipath-02, 17 October 2021, <<https://www.ietf.org/archive/id/draft-ietf-pce-multipath-02.txt>>.
- [I-D.ietf-pce-segment-routing-policy-cp]
Koldychev, M., Sivabalan, S., Barth, C., Peng, S., and H. Bidgoli, "PCEP extension to support Segment Routing Policy Candidate Paths", Work in Progress, Internet-Draft, draft-ietf-pce-segment-routing-policy-cp-05, 23 May 2021, <<https://www.ietf.org/archive/id/draft-ietf-pce-segment-routing-policy-cp-05.txt>>.
- [I-D.ietf-spring-segment-routing-policy]
Filsfils, C., Talaulikar, K., Voyer, D., Bogdanov, A., and P. Mattes, "Segment Routing Policy Architecture", Work in Progress, Internet-Draft, draft-ietf-spring-segment-routing-policy-13, 28 May 2021, <<https://www.ietf.org/archive/id/draft-ietf-spring-segment-routing-policy-13.txt>>.
- [RFC3063] Ohba, Y., Katsube, Y., Rosen, E., and P. Doolan, "MPLS Loop Prevention Mechanism", RFC 3063, DOI 10.17487/RFC3063, February 2001, <<https://www.rfc-editor.org/info/rfc3063>>.
- [RFC3630] Katz, D., Kompella, K., and D. Yeung, "Traffic Engineering (TE) Extensions to OSPF Version 2", RFC 3630, DOI 10.17487/RFC3630, September 2003, <<https://www.rfc-editor.org/info/rfc3630>>.

- [RFC5305] Li, T. and H. Smit, "IS-IS Extensions for Traffic Engineering", RFC 5305, DOI 10.17487/RFC5305, October 2008, <<https://www.rfc-editor.org/info/rfc5305>>.
- [RFC5329] Ishiguro, K., Manral, V., Davey, A., and A. Lindem, Ed., "Traffic Engineering Extensions to OSPF Version 3", RFC 5329, DOI 10.17487/RFC5329, September 2008, <<https://www.rfc-editor.org/info/rfc5329>>.
- [RFC7308] Osborne, E., "Extended Administrative Groups in MPLS Traffic Engineering (MPLS-TE)", RFC 7308, DOI 10.17487/RFC7308, July 2014, <<https://www.rfc-editor.org/info/rfc7308>>.

Authors' Addresses

Balaji Rajagopalan
Juniper Networks

Email: balajir@juniper.net

Vishnu Pavan Beeram
Juniper Networks

Email: vbeeram@juniper.net

Shaofu Peng
ZTE Corporation

Email: peng.shaofu@zte.com.cn

Quan Xiong
ZTE Corporation

Email: xiong.quan@zte.com.cn

Mike Koldychev
Cisco Systems Inc.

Email: mkoldych@cisco.com

Gyan Mishra
Verizon Communications Inc.

Email: gyan.s.mishra@verizon.com

PCE Working Group
Internet-Draft
Intended status: Standards Track
Expires: 18 May 2022

B. Rajagopalan
V. Beeram
Juniper Networks
S. Peng
Q. Xiong
ZTE Corporation
M. Koldychev
Cisco Systems Inc.
G. Mishra
Verizon Communications Inc.
14 November 2021

Path Computation Element Protocol(PCEP) Extension for Color
draft-rajagopalan-pce-pcep-color-01

Abstract

Color is a 32-bit numerical attribute that is used to associate a Traffic Engineering (TE) tunnel or policy with an intent or objective (e.g. low latency). This document specifies an extension to Path Computation Element Protocol (PCEP) to carry the color attribute.

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 18 May 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. Use case: RSVP-TE Color	3
3. Protocol Operation	3
4. TLV Format	4
5. Security Considerations	5
6. IANA Considerations	5
6.1. PCEP TLV Type Indicator	5
6.2. STATEFUL-PCE-CAPABILITY TLV Flag Field	5
6.3. LSP-ERROR-CODE TLV Error Code Field	5
7. Acknowledgments	6
8. References	6
8.1. Normative References	6
8.2. Informative References	7
Authors' Addresses	8

1. Introduction

A Traffic Engineering (TE) tunnel or policy can be associated with an intent or objective (e.g. low latency) by marking it with a color. This color attribute is used as a guiding criterion for mapping services onto the TE tunnel or policy ([RFC9012]). The term color used in this document is NOT to be interpreted as the 'thread color' specified in [RFC3063] or the 'resource color' (or 'link color') specified in [RFC3630], [RFC5329], [RFC5305] and [RFC7308].

Color is part of the tuple that identifies a Segment Routing (SR) policy ([I-D.ietf-spring-segment-routing-policy]) and is included in the Path Computation Element Protocol (PCEP) extensions defined for carrying the SR policy identifiers ([I-D.ietf-pce-segment-routing-policy-cp]). The color encoding specified in SR policy identifier cannot be reused for other types of path setup.

This document introduces a generic optional PCEP TLV called the Color TLV to carry the color attribute and discusses its usage with RSVP-TE Label Switched Paths (LSPs).

In addition to catering to the use-case discussed in this document, the Color TLV can also be used to reference SR Composite Candidate Paths as specified in ([I-D.ietf-pce-multipath]). An implementation MAY also provide a local policy option to use this TLV to reference a set of path constraints and optimization objectives.

2. Use case: RSVP-TE Color

The color attribute can be used as one of the guiding criteria in selecting the RSVP-TE LSP as a next hop for service prefixes. While the specific details of how the service prefixes are associated with the appropriate RSVP-TE LSPs are outside the scope of this specification, the envisioned high level usage of the color attribute is as follows.

The service prefixes are marked with some indication of the type of underlay they need. The underlay LSPs 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 LSPs.

As an example, for a BGP-based service, the originating PE could attach some community, e.g. the Color Extended Community [RFC9012] with the service route. A receiving PE could use locally configured policies to associate service routes carrying Color Extended Community 'X' with underlay RSVP-TE LSPs of color 'Y'.

BGP Color Extended Community is commonly used to perform service mapping, although this specification does not mandate its usage.

The procedure discussed for service mapping in this section can be applied to any underlay path setup type.

3. Protocol Operation

The STATEFUL-PCE-CAPABILITY negotiation message is enhanced to carry the color capability, which allows PCC (Path Computation Client) and PCE (Path Computation Element) 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

- * SHOULD NOT report color to the PCE.
- * MUST NOT override the local color, if it is configured, based on any messages coming from the PCE.

Section 4 defines the format of the color TLV. The placement of the TLV depends on the purpose for which it is used. For RSVP's service mapping use case discussed in this document, the color TLV is carried 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" (TBA3) in LSP State Report message.

When LSPs that belong to the same TE tunnel are with in the same Path Protection Association Group [RFC8745], the color is attached only to the primary LSP. If PCC receives color TLV for a secondary LSP, it SHOULD respond with an error code of 4 (Unacceptable Parameters).

4. TLV Format

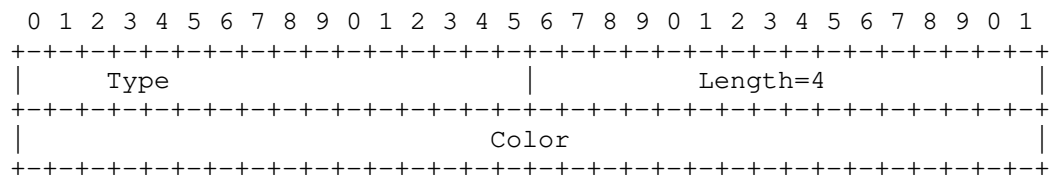


Figure 1: Color TLV

Type has the value TBA1. 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 (TBA2): A PCE/PCC that supports color capability must turn on this bit.

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

6.1. PCEP TLV Type Indicator

IANA is requested to allocate a new value in the "PCEP TLV Type Indicators" sub-registry of the PCEP Numbers registry as follows:

Value	Description	Reference
TBA1	Color	This document

6.2. STATEFUL-PCE-CAPABILITY TLV Flag Field

IANA is requested to allocate a new bit value in the "STATEFUL-PCE-CAPABILITY TLV Flag Field" sub-registry of the PCEP Numbers registry as follows:

Value	Description	Reference
TBA2	COLOR-CAPABILITY	This document

6.3. LSP-ERROR-CODE TLV Error Code Field

IANA is requested to allocate a new error code in the "LSP-ERROR-CODE TLV Error Code Field" sub-registry of the PCEP Numbers registry as follows:

Value	Meaning	Reference
TBA3	Unsupported Color	This document

7. Acknowledgments

The authors would like to thank Kaliraj Vairavakkalai, Colby Barth, Natrajan Venkataraman and Tarek Saad for their review and suggestions.

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>>.
- [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>>.

- [RFC8745] Ananthakrishnan, H., Sivabalan, S., Barth, C., Minei, I., and M. Negi, "Path Computation Element Communication Protocol (PCEP) Extensions for Associating Working and Protection Label Switched Paths (LSPs) with Stateful PCE", RFC 8745, DOI 10.17487/RFC8745, March 2020, <<https://www.rfc-editor.org/info/rfc8745>>.
- [RFC9012] Patel, K., Van de Velde, G., Sangli, S., and J. Scudder, "The BGP Tunnel Encapsulation Attribute", RFC 9012, DOI 10.17487/RFC9012, April 2021, <<https://www.rfc-editor.org/info/rfc9012>>.

8.2. Informative References

- [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", Work in Progress, Internet-Draft, draft-ietf-pce-multipath-03, 25 October 2021, <<https://www.ietf.org/archive/id/draft-ietf-pce-multipath-03.txt>>.
- [I-D.ietf-pce-segment-routing-policy-cp]
Koldychev, M., Sivabalan, S., Barth, C., Peng, S., and H. Bidgoli, "PCEP extension to support Segment Routing Policy Candidate Paths", Work in Progress, Internet-Draft, draft-ietf-pce-segment-routing-policy-cp-06, 22 October 2021, <<https://www.ietf.org/archive/id/draft-ietf-pce-segment-routing-policy-cp-06.txt>>.
- [I-D.ietf-spring-segment-routing-policy]
Filsfils, C., Talaulikar, K., Voyer, D., Bogdanov, A., and P. Mattes, "Segment Routing Policy Architecture", Work in Progress, Internet-Draft, draft-ietf-spring-segment-routing-policy-14, 25 October 2021, <<https://www.ietf.org/archive/id/draft-ietf-spring-segment-routing-policy-14.txt>>.
- [RFC3063] Ohba, Y., Katsube, Y., Rosen, E., and P. Doolan, "MPLS Loop Prevention Mechanism", RFC 3063, DOI 10.17487/RFC3063, February 2001, <<https://www.rfc-editor.org/info/rfc3063>>.
- [RFC3630] Katz, D., Kompella, K., and D. Yeung, "Traffic Engineering (TE) Extensions to OSPF Version 2", RFC 3630, DOI 10.17487/RFC3630, September 2003, <<https://www.rfc-editor.org/info/rfc3630>>.

- [RFC5305] Li, T. and H. Smit, "IS-IS Extensions for Traffic Engineering", RFC 5305, DOI 10.17487/RFC5305, October 2008, <<https://www.rfc-editor.org/info/rfc5305>>.
- [RFC5329] Ishiguro, K., Manral, V., Davey, A., and A. Lindem, Ed., "Traffic Engineering Extensions to OSPF Version 3", RFC 5329, DOI 10.17487/RFC5329, September 2008, <<https://www.rfc-editor.org/info/rfc5329>>.
- [RFC7308] Osborne, E., "Extended Administrative Groups in MPLS Traffic Engineering (MPLS-TE)", RFC 7308, DOI 10.17487/RFC7308, July 2014, <<https://www.rfc-editor.org/info/rfc7308>>.

Authors' Addresses

Balaji Rajagopalan
Juniper Networks

Email: balajir@juniper.net

Vishnu Pavan Beeram
Juniper Networks

Email: vbeeram@juniper.net

Shaofu Peng
ZTE Corporation

Email: peng.shaoфу@zte.com.cn

Quan Xiong
ZTE Corporation

Email: xiong.quan@zte.com.cn

Mike Koldychev
Cisco Systems Inc.

Email: mkoldych@cisco.com

Gyan Mishra
Verizon Communications Inc.

Email: gyan.s.mishra@verizon.com

PCE Working Group
Internet-Draft
Intended status: Standards Track
Expires: April 23, 2022

Y. Wang
A. Wang
China Telecom
October 20, 2021

PCEP Procedures and Extension for VLAN-based Traffic Forwarding
draft-wang-pce-vlan-based-traffic-forwarding-01

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 April 23, 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. Procedures for VLAN-based Traffic Forwarding	3
5. Capability Advertisement	4
6. PCEP message	5
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	13
8.1. VLAN forwarding CCI Object	13
8.2. Address TLVs	15
8.3. VLAN crossing CCI Object	15
9. Deployment Considerations	16
10. Security Considerations	16
11. IANA Considerations	16
11.1. Path Setup Type Registry	16
11.2. PCECC-CAPABILITY sub-TLV's Flag field	17
11.3. PCEP Object Types	17
11.4. PCEP-Error Object	17
12. Acknowledgement	18
13. Normative References	18
Authors' Addresses	19

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. 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 [I-D.ietf-pce-pcep-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 specifies 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 [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:

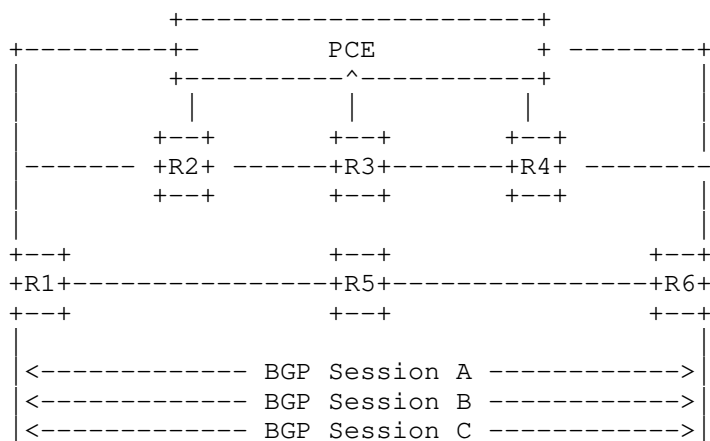


Figure 1: BGP Session Establishment Procedures

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,:

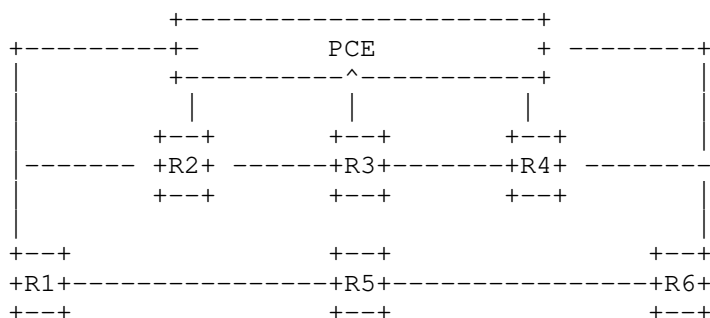


Figure 2: BGP Prefix Advertisement Procedures

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 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 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.

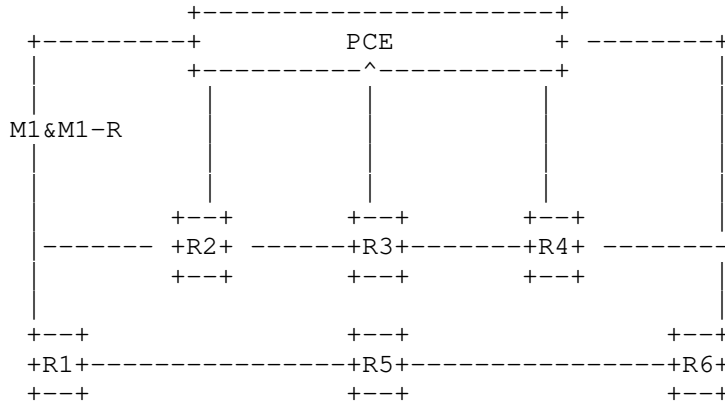


Figure 3: VLAN-Based forwarding info Advertisement Procedures for Ingress PCC

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 M1-R	PCE/R1	PCInitiate PCRpt	CC-ID=X1 VLAN Forwarding CCI Object (Peer_IP=R6_A, Interface_Address=INF1, VLAN_ID=VLAN_R1_R2)

7.3.2. VLAN-Based crossing info Advertisement Procedures

The detail procedures for VLAN-Based crossing info advertisement contained in the VLAN crossing CCI Object 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 PCC will be replaced by a new out-VLAN tag of the previous 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 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.

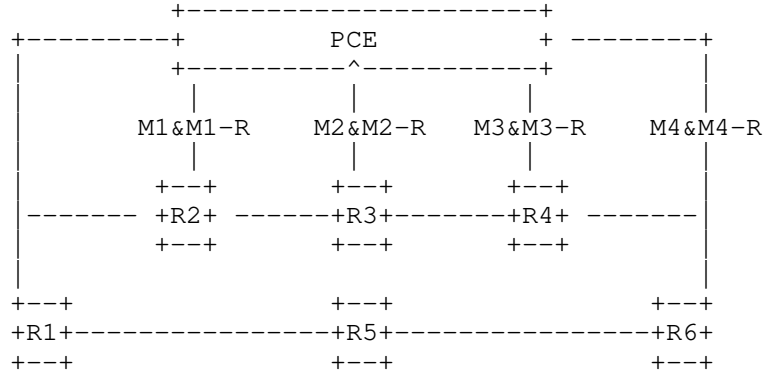


Figure 4: VLAN-Based crossing info Advertisement Procedures
for transit PCC and egress PCC

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) (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 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 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 VLAN crossing CCI Object (IN) (O=0, Interface_Address=INF1, IN_VLAN_ID=VLAN_R4_R6) VLAN crossing CCI Object (OUT) (O=1, Interface_Address=INF2, OUT_VLAN_ID=0)

8. New PCEP Objects

The Central Control Instructions (CCI) Object is used by the PCE to specify the forwarding instructions is defined in [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.

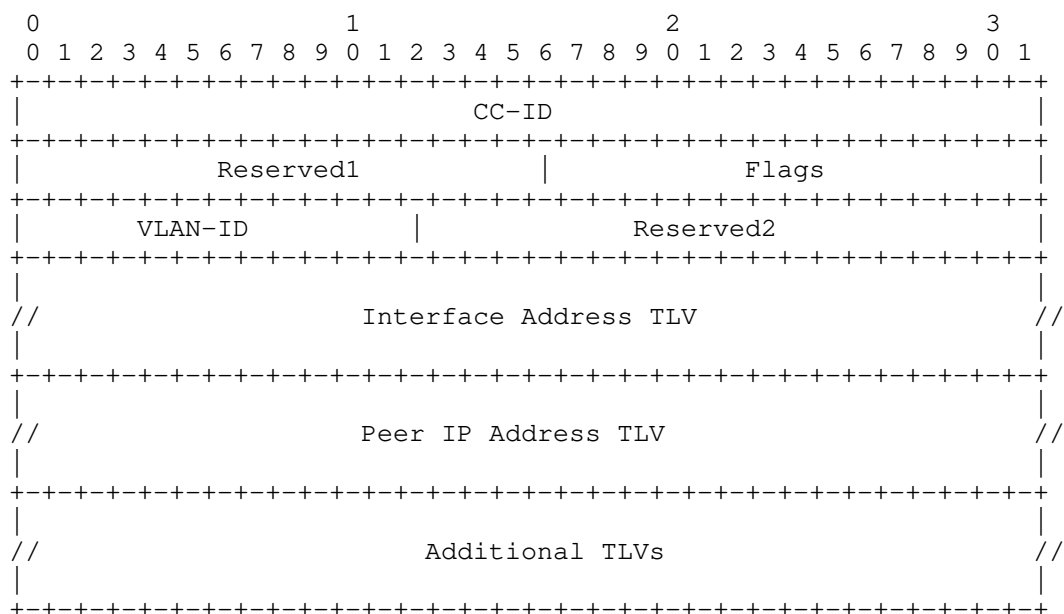


Figure 5: VLAN Forwarding CCI Object

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.

Interface Address TLV [RFC8779] MUST be included in this CCI Object-Type TBD8 to specify the interface which will set up the vlan defined in the VLAN Forwarding CCI Object.

The Peer IP Address TLV[RFC8779]MUST be included in this CCI Object-Type TBD8 to identify the end to end TE path in VLAN-based traffic forwarding network and MUST be unique.

8.2. Address TLVs

[RFC8779] defines IPV4-ADDRESS, IPV6-ADDRESS, and UNNUMBERED-ENDPOINT TLVs for the use of Generalized Endpoint. The same TLVs can also be used in the CCI object to find the Peer address that matches egress PCC and further identify the packet to be guaranteed. If the PCC is not able to resolve the peer information or can not find the corresponding ingress device, it MUST reject the CCI and respond with a PCErr message with Error-Type = TBD6 ("VLAN-based forwarding failure") and Error Value = TBD9 ("Invalid egress PCC information").

8.3. VLAN crossing CCI Object

The VLAN crossing CCI object is defined to control the transmission-path of the packet by VLAN-ID. This new type of CCI Object can be carried within a PCInitiate message sent by the PCE to the transit PCC and the egress PCC in the VLAN-based traffic forwarding scenarios.

CCI Object-Class is 44.

CCI Object-Type is TBD10 for VLAN crossing info in the native IP network.

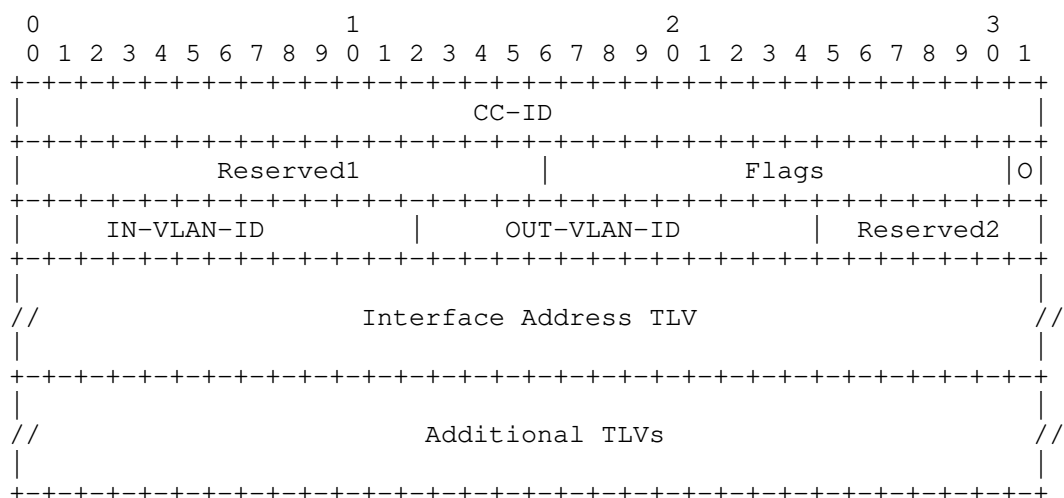


Figure 6: VLAN Crossing CCI Object

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, 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.

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.

Interface Address TLV [RFC8779] MUST be included in this CCI Object-Type TBD8 to specify the interface which will set up the vlan defined in the VLAN Forwarding CCI Object.

9. 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,
"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-16 (work in progress), August
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.

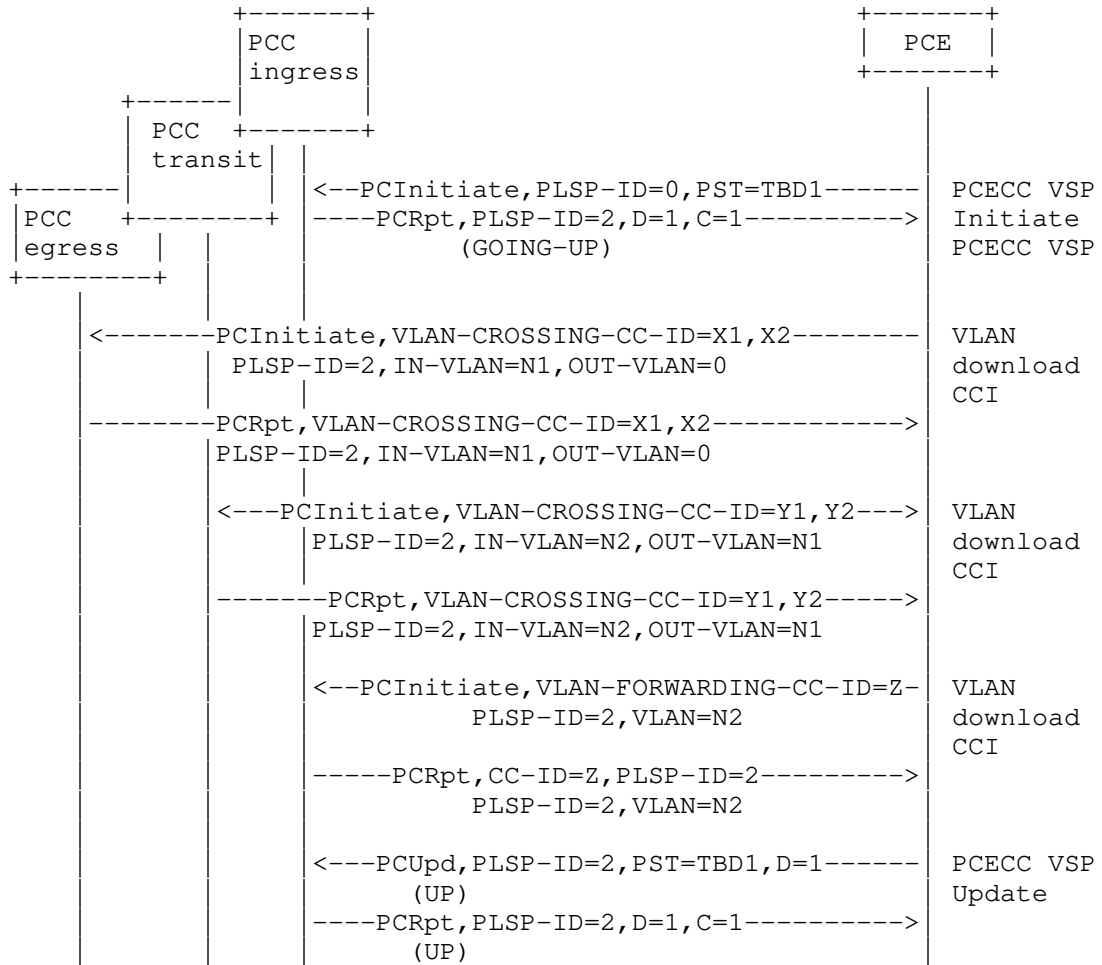
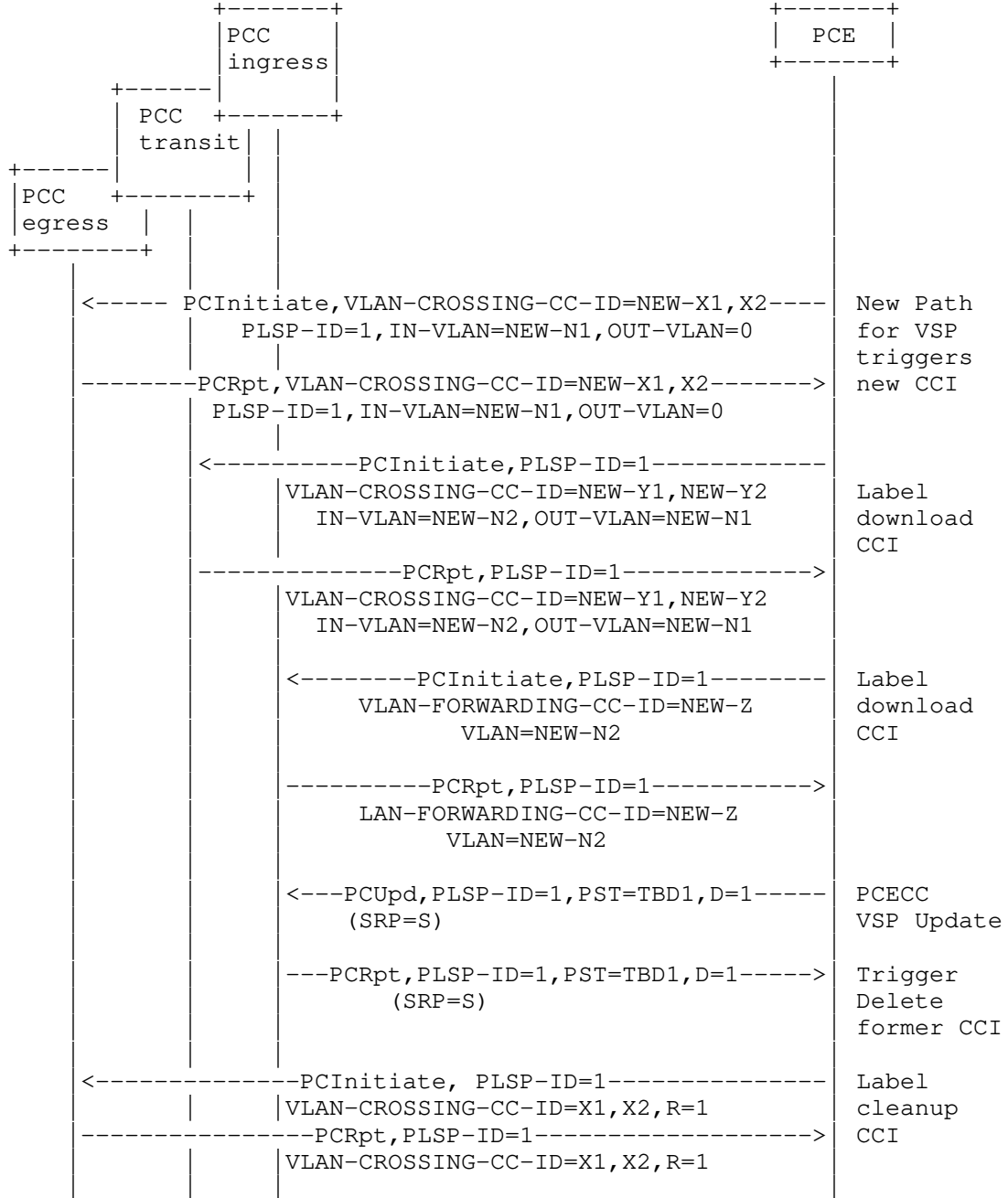


Figure 1: PCE-Initiated PCECC VSP

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.



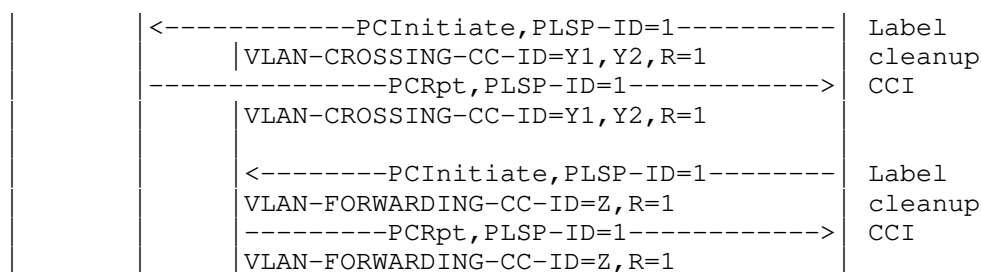


Figure 2: PCECC VSP Update

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:

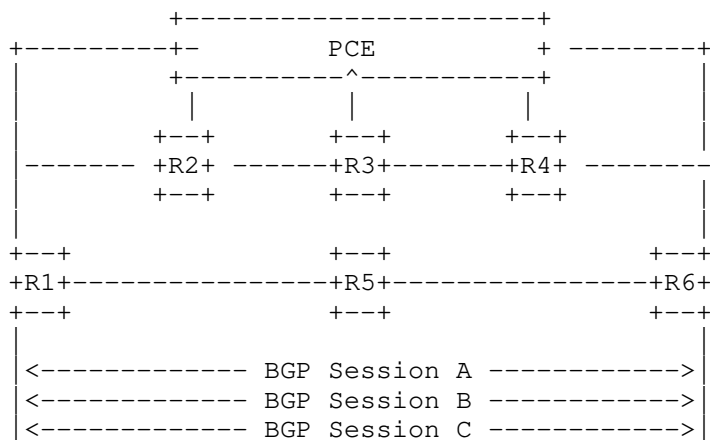


Figure 3: BGP Session Establishment Procedures

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.

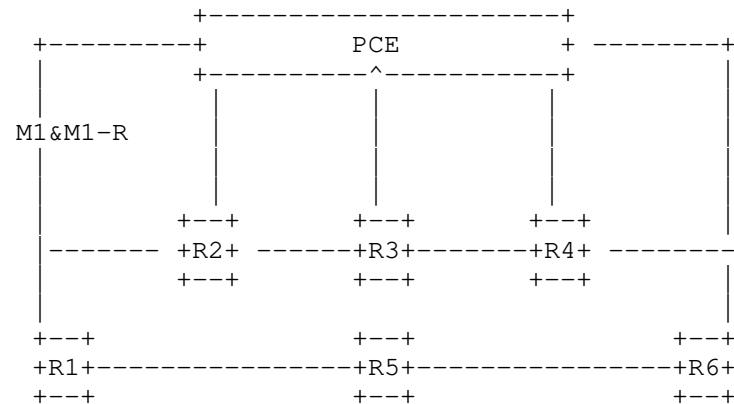


Figure 4: VLAN-Based forwarding info Advertisement Procedures for Ingress PCC

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.

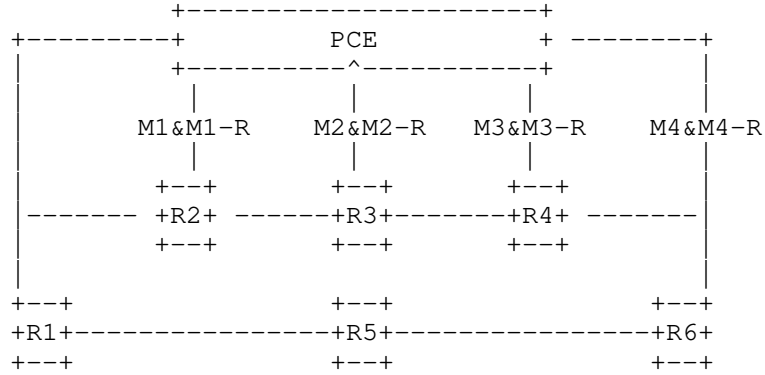


Figure 5: VLAN-Based crossing info Advertisement Procedures
for transit PCC and egress PCC

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.

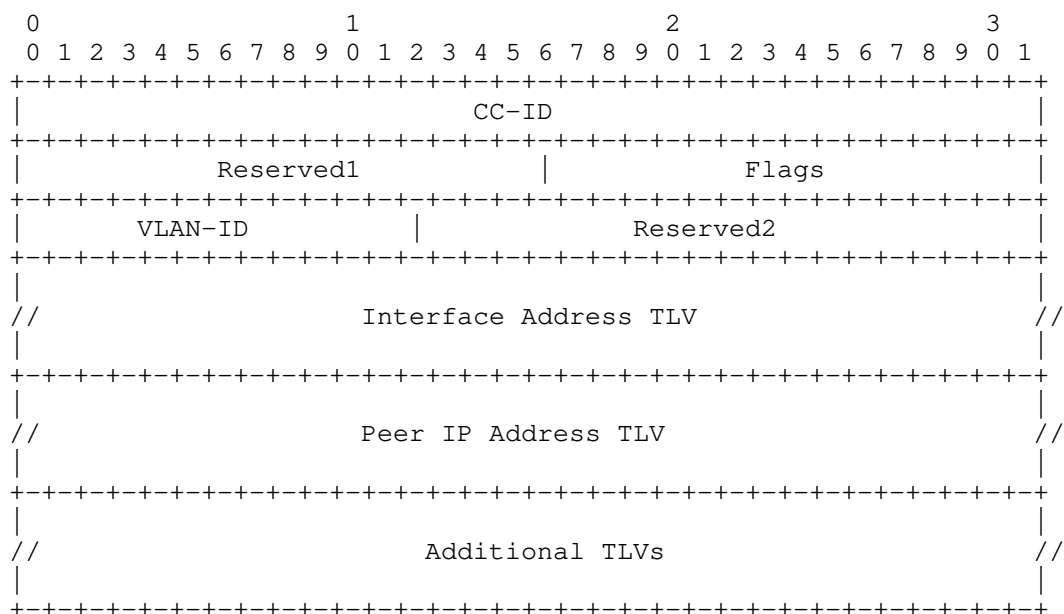


Figure 6: VLAN Forwarding CCI Object

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.

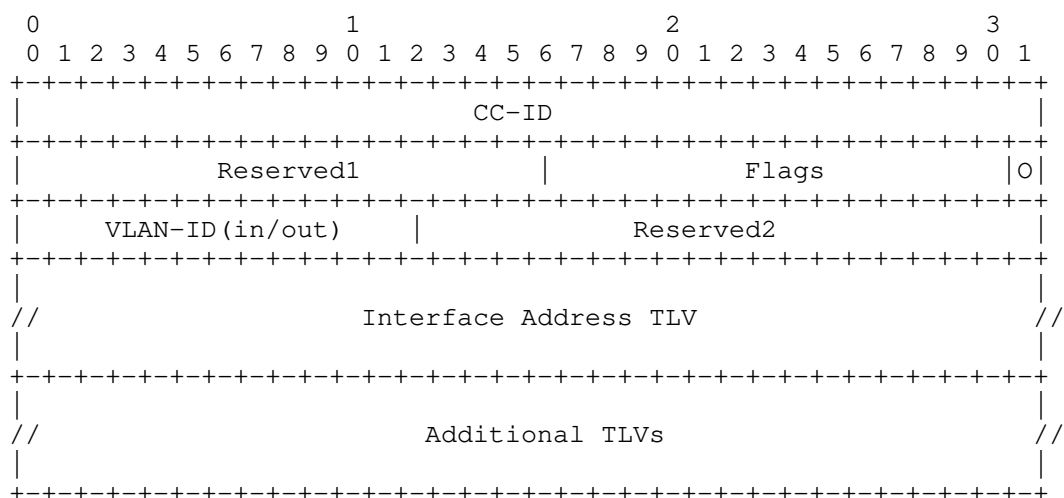


Figure 7: VLAN Crossing CCI Object

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

PCE
Internet-Draft
Intended status: Standards Track
Expires: 11 April 2022

Q. Xiong
S. Peng
ZTE Corporation
V. Beeram
T. Saad
Juniper Networks
M. Koldychev
Cisco Systems
8 October 2021

PCEP Extensions for Topology Filter
draft-xpbs-pce-topology-filter-01

Abstract

This document proposes a set of extensions for PCEP to support the topology filter during path computation.

Status of This Memo

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

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at <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 11 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	2
1.1. Terminology	3
1.2. Requirements Language	3
2. Topology Filter	3
2.1. Topology Reference	3
2.2. Filters	4
3. PCEP Extensions	4
3.1. TOPOLOGY Object	4
3.1.1. Source Protocol TLV	5
3.1.2. Multi-topology TLV	5
3.1.3. Area TLV	6
3.1.4. Algorithm TLV	7
3.2. IRO Object	7
3.2.1. Link ID	7
3.2.2. Admin Group	7
3.2.3. Source Protocol	8
3.3. XRO Object	9
3.4. Procedures	9
4. Acknowledgements	9
5. IANA Considerations	9
5.1. TOPOLOGY Object	9
5.2. IRO and XRO Object	10
6. Security Considerations	10
7. References	10
7.1. Normative References	10
Authors' Addresses	11

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.

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 sub-topology during path computation. The PCE MUST take the topology constraint into consideration during path computation.

The sub-topology may be considered as a TE topology or a specific IGP domain. As defined in draft-bestbar-teas-yang-topology-filter, a topology filter is a data construct that can be applied on either a native topology or a user specified topology. The topology filter can be viewed as a set of filtering rules to construct the sub-topology. The topology filter specifies the topology reference or a set of include-any, include-all and exclude filtering rules.

This document proposes a set of extensions for PCEP to support the topology filter during path computation.

1.1. Terminology

The terminology is defined as [RFC5440], [RFC7752] and [RFC8795].

1.2. Requirements Language

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

2. Topology Filter

As defined in draft-bestbar-teas-yang-topology-filter, a topology filter is a data construct that can be applied on either a native topology or a user specified topology. The topology filter can be viewed as a set of filtering rules to construct the sub-topology. The topology filter specifies the topology reference or a set of include-any, include-all and exclude filtering rules.

2.1. Topology Reference

The topology reference indicates the topology on which the existing referenced filtering rules need to be applied. The referenced topology could be a predefined TE topology or a specific IGP domain.

As defined in RFC7752, the IGP domain has a unique IGP representation by using the combination of Area-ID, Router-ID, Protocol-ID, Multi-Topology ID, and Instance-ID. This document defines TOPOLOGY object and new TLVs for the topology filter such as Source Protocol TLV, Multi-Topology ID, Area-ID and Algorithm TLV.

2.2. Filters

The topology filters carries a list of filters. Each filter specifies a set of include-any, include-all and exclude filtering rules that can be applied on the native topology. The filtering rules specify the a set of constraints on the topology, that are to be used to compute path at PCE. This document proposes a set of extensions for IRO and XRO object and defines new subobjects such as Link ID, Link affinity and Source Protocol.

3. PCEP Extensions

3.1. TOPOLOGY Object

This document defines a new TOPOLOGY object to carry the topology filter.

The TOPOLOGY object is optional and specifies the sub-topology to be taken into account by the PCE during path computation. The TOPOLOGY object can be carried within a PCReq message, or a PCRep message in case of unsuccessful path computation.

TOPOLOGY Object-Class is TBD1.

TOPOLOGY Object-Type is TBD2.

The format of the TOPOLOGY object body is:

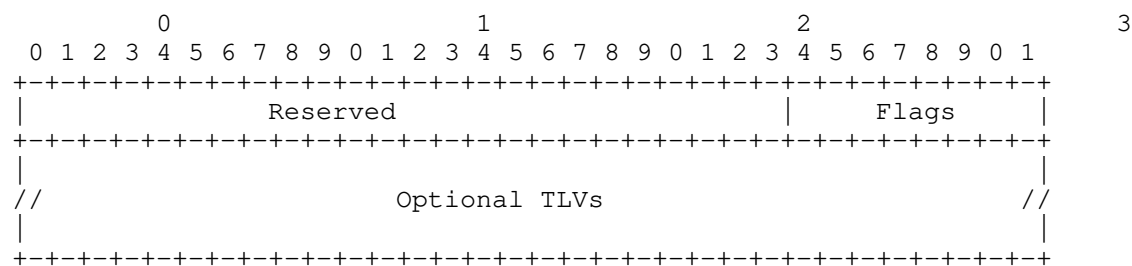


Figure 1: TOPOLOGY Body Object Format

Reserved (24 bits): This field MUST be set to zero on transmission and MUST be ignored on receipt.

Flags (8 bits): No flags are currently defined. Unassigned flags MUST be set to zero on transmission and MUST be ignored on receipt.

The format of optional TLVs is defined in RFC5440 and may be used to carry topology filter information as defined in section.

3.1.1. Source Protocol TLV

The Source Protocol TLV is optional and is defined to carry the protocol ID and Instance ID.

The format of the Source Protocol TLV is:

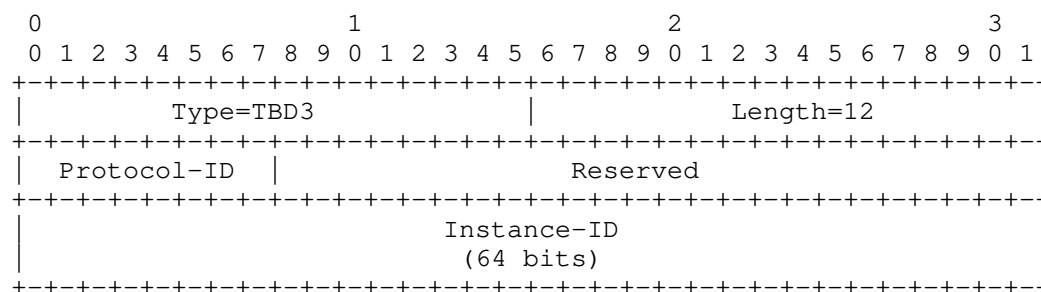


Figure 2: Source Protocol TLV

The code point for the TLV type is TBD3. The TLV length is 12 octets.

Protocol-ID (8 bits): defined in [RFC7752] section 3.2. IS-IS [RFC8202] and OSPF [RFC6549] MAY run multiple routing protocol instances identified by the Protocol-ID over the same link.

Reserved (24 bits): This field MUST be set to zero on transmission and MUST be ignored on receipt.

Instance-ID (64 bits): defined in [RFC7752] section 3.2.

3.1.2. Multi-topology TLV

The Multi-topology TLV is optional and is defined to carry the multi-topology ID.

The format of the Multi-topology TLV is :

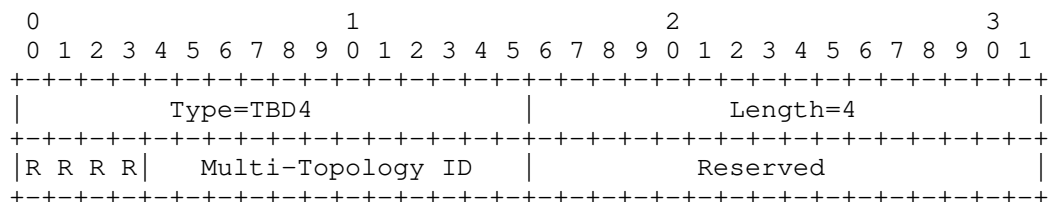


Figure 3: Multi-topology TLV

The code point for the sub-TLV type is TBD4. The sub-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]. As defined in section 3.2.1.5 of [RFC7752], 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.

3.1.3. Area TLV

The Area TLV is optional and is defined to carry the Area ID.

The format of the Area TLV is :

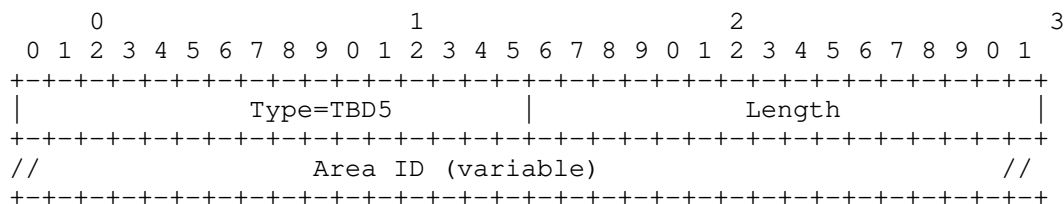


Figure 4: Area TLV

The code point for the TLV type is TBD3. The TLV length is variable.

Area-ID: Area identifier as defined in RFC7752.

3.1.4. Algorithm TLV

The Algorithm TLV is optional and is defined to carry the Algorithm ID.

The Algorithm TLV MAY be inserted so as to provide the Flex-algo plane information for the computed path. The format of the TLV is defined in draft-tokar-pce-sid-algo-04 section 3.4.

3.2. IRO Object

As per [RFC5440], IRO can be used to specify that the computed path needs to traverse a set of specified network elements or abstract nodes. This document proposed a set of extensions for topology filter.

3.2.1. Link ID

The Link ID is used to include the link that is used during the path calculation.

The Link ID subobject is defined:

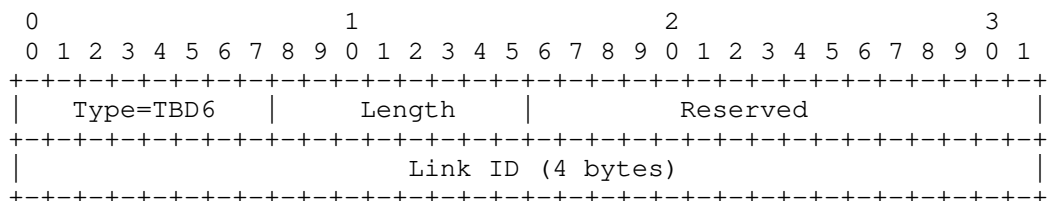


Figure 5: Link ID subobject in IRO

The code point for the TLV type is TBD6. The TLV length is 12 octets.

Link ID (32bits): defined in IS-IS RFC5307 and OSPF RFC3630.

3.2.2. Admin Group

The Admin Group is used to include the links that is used during the path calculation.

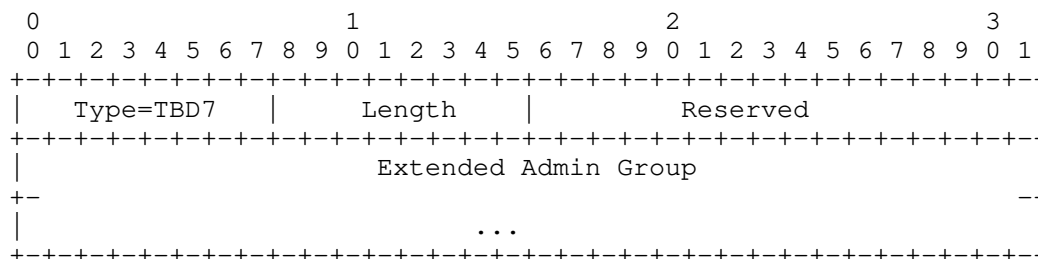


Figure 6: Admin Group subobject in IRO

The code point for the TLV type is TBD7. The TLV length is variable.

Extended Administrative Group: Extended Administrative Group as defined in [RFC7308].

3.2.3. Source Protocol

The format of the Source Protocol subobject is:

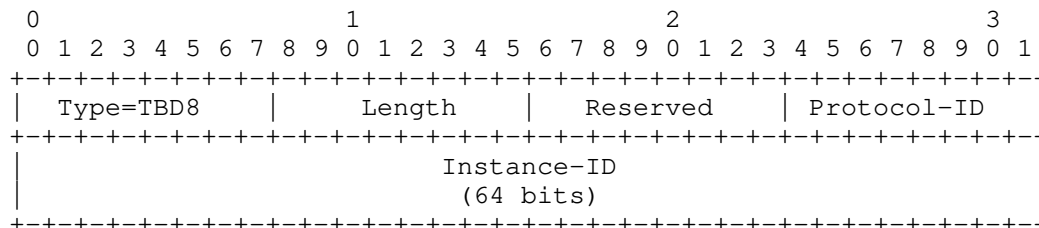


Figure 7: Source Protocol subobject in IRO

The code point for the TLV type is TBD8. The TLV length is 12 octets.

Protocol-ID (8 bits): defined in [RFC7752] section 3.2. IS-IS [RFC8202] and OSPF [RFC6549] MAY run multiple routing protocol instances identified by the Protocol-ID over the same link.

Reserved (24 bits): This field MUST be set to zero on transmission and MUST be ignored on receipt.

Instance-ID (64 bits): defined in [RFC7752] section 3.2.

3.3. XRO Object

As per [RFC5521], XRO is an optional object used to specify exclusion of certain abstract nodes or resources from the whole path. This document proposed a set of extensions for topology filter.

The XRO is made of sub-objects identical to the ones defined in IRO, where the XRO sub-object type is identical to the sub-object type defined in this documents.

The following sub-object types are supported.

Type Sub-object

TBD6 Link ID

TBD7 Admin Group

TBD8 Source Protocol

3.4. Procedures

A PCC MAY insert a TOPOLOGY object to indicate the sub-topology of an IGP domain that MUST be considered by the PCE. The PCE will perform path computation based on the sub-topology identified by the topology filter rules that can be applied on either the native topology or a user specified topology. The absence of the TLVs related topology reference indicates that the filtering rules are to be applied on the native topology.

4. Acknowledgements

TBA

5. IANA Considerations

5.1. TOPOLOGY Object

IANA is requested to make allocations for Topology Object from the registry, as follows:

TOPOLOGY Object-Class is TBD1.

TOPOLOGY Object-Type is TBD2.

The TLVs for Topology Object is as follows:

Type	TLV	Reference
TBD3	Source Protocol TLV	[this document]
TBD4	Multi-topology TLV	[this document]
TBD5	Area TLV	[this document]

Table 1: TLVs for Topology Object

5.2. IRO and XRO Object

IANA is requested to make allocations for IRO and ERO Object from the registry, as follows:

Type	Subobject	Reference
TBD6	Link ID	[this document]
TBD7	Admin Group	[this document]
TBD8	Source Protocol	[this document]

Table 2: Subobjects for IRO and XRO Object

6. Security Considerations

TBA

7. References

7.1. Normative References

- [draft-ietf-lsr-flex-algo]
 "IGP Flexible Algorithm", July 2021, <<https://www.rfc-editor.org/info/draft-ietf-lsr-flex-algo>>.
- [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] "A Path Computation Element (PCE)-Based Architecture", August 2006, <<https://www.rfc-editor.org/info/RFC4655>>.

- [RFC4915] "Multi-Topology (MT) Routing in OSPF", June 2007,
<<https://www.rfc-editor.org/info/RFC4915>>.
- [RFC5120] "M-ISIS: Multi Topology (MT) Routing in Intermediate
System to Intermediate Systems (IS-ISs)", February 2008,
<<https://www.rfc-editor.org/info/RFC5120>>.
- [RFC5440] "Path Computation Element (PCE) Communication Protocol
(PCEP)", March 2009,
<<https://www.rfc-editor.org/info/RFC5440>>.
- [RFC6549] "OSPFv2 Multi-Instance Extensions", March 2012,
<<https://www.rfc-editor.org/info/RFC6549>>.
- [RFC7752] "North-Bound Distribution of Link-State and Traffic
Engineering (TE) Information Using BGP", March 2016,
<<https://www.rfc-editor.org/info/RFC7752>>.
- [RFC8174] "Ambiguity of Uppercase vs Lowercase in RFC 2119 Key
Words", May 2017,
<<https://www.rfc-editor.org/info/RFC8174>>.
- [RFC8202] "IS-IS Multi-Instance", June 2017,
<<https://www.rfc-editor.org/info/RFC8202>>.
- [RFC8231] "Path Computation Element Communication Protocol (PCEP)
Extensions for Stateful PCE", September 2017,
<<https://www.rfc-editor.org/info/RFC8231>>.
- [RFC8795] "YANG Data Model for Traffic Engineering (TE) Topologies",
August 2020, <<https://www.rfc-editor.org/info/RFC8795>>.

Authors' Addresses

Quan Xiong
ZTE Corporation
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

Vishnu Pavan Beeram
Juniper Networks

Email: vbeeram@juniper.net

Tarek Saad
Juniper Networks

Email: tsaad@juniper.net

Mike Koldychev
Cisco Systems
Canada

Email: mkoldych@cisco.com