Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: August 25, 2021

H. Chen
M. McBride
Futurewei
A. Wang
China Telecom
G. Mishra
Verizon Inc.
Y. Liu
China Mobile
Y. Fan
Casa Systems
L. Liu
Fujitsu
X. Liu
Volta Networks

February 21, 2021

PCE for BIER-TE Path
draft-chen-pce-bier-te-path-00

Abstract

This document describes extensions to Path Computation Element (PCE) communication Protocol (PCEP) for supporting Bit Index Explicit Replication (BIER) Traffic Engineering (TE) paths.

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 August 25, 2021.

Table of Contents

1. Introduction

[I-D.ietf-bier-te-arch] introduces Bit Index Explicit Replication (BIER) Traffic/Tree Engineering (BIER-TE). It is an architecture for per-packet stateless explicit point to multipoint (P2MP) multicast path/tree and based on the BIER architecture defined in [RFC8279].

A Bit-Forwarding Ingress Router (BFIR) in a BIER-TE domain receives the information or instructions from a controller such as a stateful PCE about which multicast flows/packets are mapped to which P2MP paths. The multicast flows/packets are indicated by multicast and source addresses. The paths are represented by BitPositions or say BitStrings. After receiving the information or instructions, the ingress node/router encapsulates the multicast packets with the BitPositions for the corresponding P2MP paths, replicates and forwards the packets with the BitPositions along the P2MP paths.

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

To compute and initiate BIER-TE P2MP paths, the stateful PCE needs to be extended. For a BIER-TE P2MP path, some new state information will be stored and maintained, which includes the BitPositions, multicast group and multicast source for the path. The PCE gets the egresses of the path, the same multicast group and source from the egresses when each of the egresses reports to the PCE that it receives a multicast join with the multicast group and source. With this information, the PCE finds an ingress for the path, computes the path from the ingress to the egresses that has the optimal BitPositions and satisfies the constraints, and then initiates the BIER-TE path at the ingress of the path through sending the ingress the BitPositions of the path, multicast group and source in a PCEP

message such as PCInitiate. After receiving the message, the ingress creates a forwarding entry that imports the packets with the multicast group/address and source into the BIER-TE path (i.e., encapsulates the packets with a BIER-TE header having the BitPositions of the path), and then reports the status of the path to the PCE in a PCEP message such as PCRpt.

[I-D.chen-pce-bier] describes part of the solution for this, which is mainly the BIER-ERO subobject used for P2MP paths.

This document proposes a comprehensive solution for computing and establishing BIER-TE P2MP paths.

1.1. Terminologies

The following terminologies are used in this document.

PCE: Path Computation Element

PCEP: PCE communication Protocol

PCC: Path Computation Client

CE: Customer Edge

PE: Provider Edge

BIER: Bit Index Explicit Replication.

BIER-TE: BIER Traffic/Tree Engineering.

BFR: Bit-Forwarding Router.

BFIR: Bit-Forwarding Ingress Router.

BFER: Bit-Forwarding Egress Router.

BFR-id: BFR Identifier. It is a number in the range [1,65535].

BFR-NBR: BFR Neighbor.

BFR-prefix: An IP address (either IPv4 or IPv6) of a BFR.

BIRT: Bit Index Routing Table. It is a table that maps from the BFR-id (in a particular sub-domain) of a BFER to the BFR-prefix of that BFER, and to the BFR-NBR on the path to that BFER.

BIFT: Bit Index Forwarding Table.

LSP-DB: Label Switching Path DataBase.

TED: Traffic/Tree Engineering DataBase.

2. Overview of PCE for BIER-TE

This section briefly describes PCE for BIER-TE and illustrates some details through a simple example BIER-TE topology.

2.1. Example BIER-TE Topology with PCE

An example BIER-TE topology for a BIER-TE domain with a PCE is shown in Figure 1. There are 8 nodes/BFRs A, B, C, D, E, F, G and H in the domain. Nodes/BFRs A, H, E, F and D are BFIRs (i.e., ingress nodes) or BFERs (i.e., egress nodes). There is a connection (i.e., PCE session) between the PCE and the PCC running on each of the possible ingress and egress nodes in the domain. Note that some of connections and the PCC on each node are not shown in the figure.

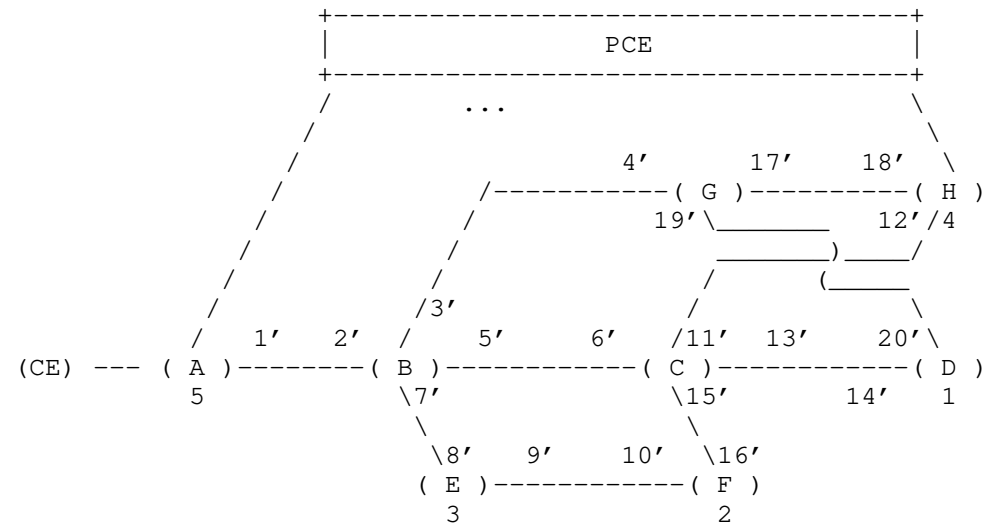


Figure 1: Example BIER-TE Topology with PCE

Nodes/BFRs D, F, E, H and A are BFERs (or BFIRs) and have local decap adjacency BitPositions 1, 2, 3, 4, and 5 respectively. For simplicity, these BPs are represented by (SI:BitString), where SI = 0 and BitString is of 8 bits. BPs 1, 2, 3, 4, and 5 are represented by 1 (0:00000001), 2 (0:00000010), 3 (0:00000100), 4 (0:00001000) and 5 (0:00010000) respectively.

The BitPositions for the forward connected adjacencies are represented by i', where i is from 1 to 20. In one option, they are encoded as (n+i), where n is a power of 2 such as 32768. For simplicity, these BitPositions are represented by (SI:BitString), where SI = (6 + (i-1)/8) and BitString is of 8 bits. BitPositions i' (i from 1 to 20) are represented by 1'(6:00000001), 2'(6:00000010), 3'(6:00000100), 4'(6:00001000), 5'(6:00010000), 6'(6:00100000), 7'(6:01000000), 8'(6:10000000), 9'(7:00000001), 10'(7:00000010), . . . , 16'(7:10000000), 17'(8:00000001), 18'(8:00000010), . . . , 20'(8:00001000).

For a link between two nodes X and Y, there are two BitPositions for two forward connected adjacencies. These two forward connected adjacency BitPositions are assigned on nodes X and Y respectively. The BitPosition assigned on X is the forward connected adjacency of Y. The BitPosition assigned on Y is the forward connected adjacency of X.

For example, for the link between nodes B and C in the figure, two forward connected adjacency BitPositions 5' and 6' are assigned to two ends of the link. BitPosition 5' is assigned on node B to B's end of the link. It is the forward connected adjacency of node C. BitPosition 6' is assigned on node C to C's end of the link. It is the forward connected adjacency of node B.

2.2. A Brief Flow of PCEP Messages for a BIER-TE Path

For a BIER-TE Path to transport the packets with a given multicast group/address and source in a BIER-TE domain, a sequence of PCEP messages are exchanged between the PCE for the domain and the PCEs for the domains containing the source, and between the PCE for the domain and the PCCs running on the BFERs/BFIRs of the domain.

Suppose that each of nodes H, D and F receives a multicast join with a same multicast group/address and source, which are MGa and MSa respectively. For simplicity, assume that the multicast source MSa is in the left domain containing the CE in Figure 1. The following is a brief flow of PCEP messages for computing and creating a BIER-TE Path to transport the packets to H, D and F.

At first, the PCC running on each of nodes H, D and F sends the PCE a PCEP message such as PCRpt. The message contains the multicast group and source (i.e., MGa and MSa), which reports to the PCE that the node receives a multicast join with MGa and MSa. Note that a PCEP message is sent to the PCE from the PCC on a node to report that the node leaves when the node receives a multicast leave with MGa and MSa.

After receiving the PCEP messages from nodes H, D and F reporting multicast join with MGa and MSa, the PCE for the domain containing these nodes determines that nodes H, D and F are the egress nodes of a BIER-TE path since they have the same multicast group and source.

Second, the PCE for the domain sends a PCEP message such as PCReq to each of the PCEs for the domains that may contain the multicast source. This message requests the PCE (that may contain the source) to find an ingress node for the BIER-TE path having egress nodes H, D and F. The message contains the multicast group and source (i.e., MGa and MSa). For example, the PCE for the BIER-TE domain sends the PCEP message to the PCE (called PCE-L) for the left domain containing CE (note that this PCE is not shown in the figure).

After receiving the PCEP message requesting to find an ingress node, the PCE (e.g., PCE-L) for the domain containing the multicast source computes the ingress node that is reachable from the source with minimum cost (e.g., ingress node A). The PCE for the domain without the source can not find any ingress node.

Third, the PCE for the domain with the source sends the PCE for the BIER-TE domain a PCEP message such as PCRep with the ingress node. The PCE for the domain without the source sends the PCE for the BIER-TE domain a PCEP message such as PCRep with NO INGRESS FOUND.

After receiving the PCEP message with the ingress node, the PCE for the BIER-TE domain computes a P2MP path from the ingress node (e.g., A) to the egress nodes (e.g., H, D and F). The path has the optimal BitPositions and satisfies the constraints. The optimal BitPositions means the BitPositions for the path has the minimum number of bit sets and the minimum bit distance.

Fourth, the PCE for the BIER-TE domain sends a PCEP message such as PCInitiate to the PCC on the ingress node (e.g., A) for the ingress to create a BIER-TE path to transport the packets for the given multicast group and source. The message contains the BitPositions for the path, the multicast group and source.

After receiving the PCEP message with the path, the PCC on the ingress (e.g., A) creates the BIER-TE path, i.e., a forwarding entry that imports the packets with the multicast group/address and source into the BIER-TE path (i.e., encapsulates the packets with a BIER-TE header having the BitPositions of the path).

And then the PCC on the ingress sends the PCE a PCEP message such as PCRpt reporting the status of the path to the PCE.

After receiving the PCEP message with the status of the path, the PCE for the domain updates the information about the path accordingly.

2.3. Procedures on Ingress

This section introduces the procedures for the ingress node of a P2MP path to get the BitPositions representing the explicit P2MP path from the ingress node to its egress nodes from the PCE.

Suppose that node A in Figure 1 wants to have an explicit P2MP path from ingress node A to egress nodes H and F. The path satisfies a set of constraints. In one case, the PCC running on ingress node A sends a request for the path to the PCE. The request contains the set of constraints, objective functions, the ingress node and the egress nodes. After receiving the request, the PCE computes an explicit P2MP path, which satisfies the constraints and is from the given ingress node to the egress nodes. While computing the path, the PCE will optimize the BitPositions of the path. That is that, for a given length of BitString, the path computed uses the minimum number of BitStrings (i.e., bit sets) and satisfies the constraints. The length is given by the value in BitStrLen field in the PCE-BIER-TE-Path-Capability sub-TLV. The PCE sends a reply with the path to the PCC. The reply contains the BitPositions representing the explicit P2MP path.

For example, assume that the explicit P2MP path computed by the PCE traverses the link/adjacency from A to B (indicated by BP 2’), the link/adjacency from B to G (indicated by BP 4’) and the link/adjacency from B to C (indicated by BP 6’), the link/adjacency from G to H (indicated by BP 18’), and the link/adjacency from C to F (indicated by BP 16’). This path is represented by {2’, 4’, 6’, 16’, 18’, 2, 4}, where BitPositions 2 and 4 indicate egress nodes F and H respectively. The reply sent to the PCC on node A by the PCE contains the path represented by {2’, 4’, 6’, 16’, 18’, 2, 4}.

In another case, a request for a P2MP path is from a user or application. After receiving the request, the PCE finds an ingress node if no ingress is given, and computes an explicit P2MP path from the ingress node to the egress nodes and sends the path to the PCC running on the ingress node.

After receiving the P2MP path, for any packet from CE to be transported by the path, such as the packet with the multicast address, the ingress node encapsulates the packet with the BitPositions representing the path and forwards the packet according to its BIFT.

For example, when ingress node A receives the path represented by BitPositions {2', 4', 6', 16', 18', 2, 4}, it encapsulates every packet from CE with the multicast address with the BitPositions and then forwards the packet along the P2MP path according to its BIFT.

A forwards the packet to B according to the forwarding entry for BP 2' in its BIFT.

After receiving the packet from A, B forwards the packet to G and C according to the forwarding entries for BPs 4' and 6' in B's BIFT respectively. The packet received by G has path {16', 18', 2, 4}. The packet received by C has path {16', 18', 2, 4}.

After receiving the packet from B, G sends the packet to H according to the forwarding entry for BP 18' in G's BIFT.

After receiving the packet from B, C sends the packet to F according to the forwarding entry for BP 16' in C's BIFT.

Egress node H of the P2MP path receives the packet with BitPosition 4. It decapsulates the packet and pass the payload of the packet to the packet's NextProto.

Egress node F of the P2MP path receives the packet with BitPosition 2. It decapsulates the packet and pass the payload of the packet to the packet's NextProto.

3. Extensions to PCEP

This section describes extensions to PCEP.

3.1. BIER-TE Path Capability

During a PCEP session establishment, PCEP Speakers (PCE or PCC) indicate their ability to support BIER-TE paths. The OPEN object in the Open message contains the PATH-SETUP-TYPE-CAPABILITY TLV, which is defined in [RFC8408]. The TLV contains a list of Path Setup Types (PSTs) and optional sub-TLVs associated with the PSTs. The sub-TLVs convey the parameters that are associated with the PSTs supported by a PCEP speaker.

This document defines a new PST value:

* PST = TBD1: Path is setup using BIER-TE.

A new sub-TLV associated with this new PST is defined, which is called PCE-BIER-TE-Path-Capability sub-TLV. The format of this new sub-TLV is illustrated in the figure below.

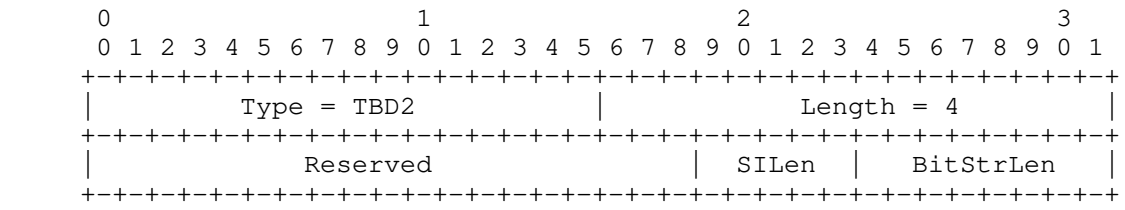


Figure 2: PCE-BIER-TE-Path-Capability sub-TLV

Type - 16 bits: TBD2 is to be assigned by IANA.

Length - 16 bits: 4 is the total length in bytes of the remainder of the TLV, excluding the Type and Length fields.

SILen (SI Length) - 5 bits: The length in bits of the SI field.

BitStrLen (Bit String Length) - 8 bits: The length in bits of the BitString field according to [RFC8296]. If k is the length of the BitString, the value of BitStrLen is log2(k)-5. For example, BitStrLen = 1 indicates k = 64, BitStrLen = 7 indicates k = 4096.

Reserved - 19 bits: MUST be set to zero by the sender and MUST be ignored by the receiver.

A PCEP speaker supporting BIER-TE paths includes the new PST and sub-TLV in the PATH-SETUP-TYPE-CAPABILITY TLV.

3.2. Extensions to SRP

For a PCEP message, when it is used for a BIER-TE path, the SRP (Stateful PCE Request Parameters) object in the message MUST include the PATH-SETUP-TYPE TLV defined in [RFC8408]. The TLV MUST contain the PST = TBD1 for path setup using BIER-TE.

Three contiguous bits in SRP Object Flag Field are defined to indicate one of the assistant operations for a BIER-TE path. This three bits field is called AOP (Assistant Operations). In addition, a new TLV, called Multicast Traffic Description TLV or Multicast Traffic TLV for short, is defined.

3.2.1. SRP Object Flag Field

The three bits for AOP are bits 27 to 29 (the exact bits to be assigned by IANA) in the SRP Object Flag Field. The values of AOP are defined as follows:

```
AOP Value    Meaning (Assistant Operation)
0x001 (J):  Join with Multicast Group and Source
0x010 (L):  Leave from Multicast Group and Source
0x011 (I):  Ingress node computation
```

The value of AOP indicates one of the three operations above. When any of the other values is received, an error MUST be reported.

When the PCC running on an edge node of a BIER-TE domain sends the PCE for the domain a PCEP message such as PCRpt to report that the edge node receives a multicast join, the message MUST include a SRP object with AOP == 0x001 (J).

When the PCC running on an edge node of a BIER-TE domain sends the PCE for the domain a PCEP message such as PCRpt to report that the edge node receives a multicast leave, the message MUST include a SRP object with AOP == 0x010 (L).

When the PCE for the domain sends a PCEP message such as PCReq to another PCE for requesting to find an ingress node for a BIER-TE path, the message MUST include a SRP object with AOP == 0x011 (I).

3.2.2. Multicast Traffic TLV

For a PCE-Initiated BIER-TE path, when a PCE sends a PCC a message such as PCInitiate message to create a BIER-TE path in a BIER-TE domain, the message MUST contain the Multicast Traffic TLV in the SRP object.

When the PCC running on an edge node of a BIER-TE domain sends the PCE for the domain a PCEP message to report that the edge node receives a multicast join or leave with a multicast group/address and source, the message MUST contain the Multicast Traffic TLV in the SRP object.

When the PCE for a BIER-TE domain sends another PCE a PCEP message to request for finding an ingress node of a BIER-TE path, the message MUST contain the Multicast Traffic TLV in the SRP object.

The format of the Multicast Traffic TLV is illustrated below.

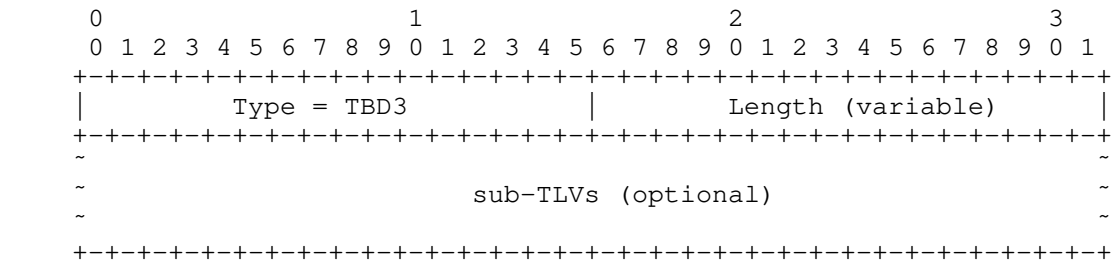


Figure 3: Multicast Traffic Description sub-TLV

Type: TBD3 is to be assigned by IANA.

Length: Variable.

Two groups of sub-TLVs are defined. One group is for IPv4, and includes IPv4 multicast group address prefix sub-TLV and IPv4 multicast source address prefix sub-TLV. The other group is for IPv6, and includes IPv6 multicast group address prefix sub-TLV and IPv6 multicast source address prefix sub-TLV.

A Multicast Traffic Description TLV MUST contain one multicast group address prefix sub-TLV, either an IPv4 or IPv6 multicast group address prefix sub-TLV. When the TLV contains an IPv4 or IPv6 multicast group address prefix sub-TLV, it may include an IPv4 or IPv6 multicast source address prefix sub-TLV respectively.

An IPv4 or IPv6 multicast group address prefix sub-TLV describes the traffic (or the packets) to be imported into the BIER-TE path/tunnel. It is an IPv4 or IPv6 multicast group address prefix. The traffic (or the packets) with the IPv4 or IPv6 multicast group address matching the prefix will be transported by the BIER-TE path/tunnel.

The formats of an IPv4 and IPv6 multicast group address prefix sub-TLV are illustrated below.

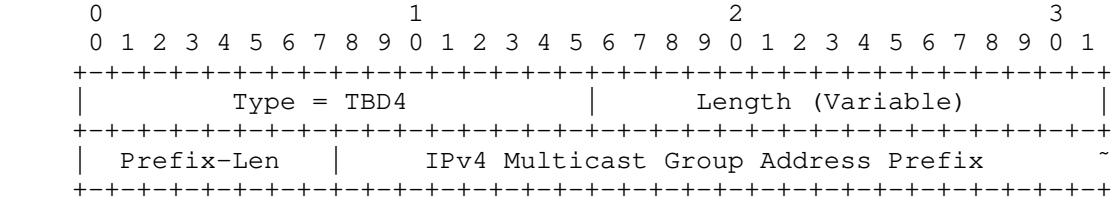


Figure 4: IPv4 Multicast Group Address Prefix sub-TLV

Type: TBD4 is to be assigned by IANA.

Length: Variable.

Prefix-Len: Indicates the length in bits of the IPv4 Multicast Group Address Prefix.

IPv4 Multicast Group Address Prefix: Indicates an IPv4 multicast group address prefix.

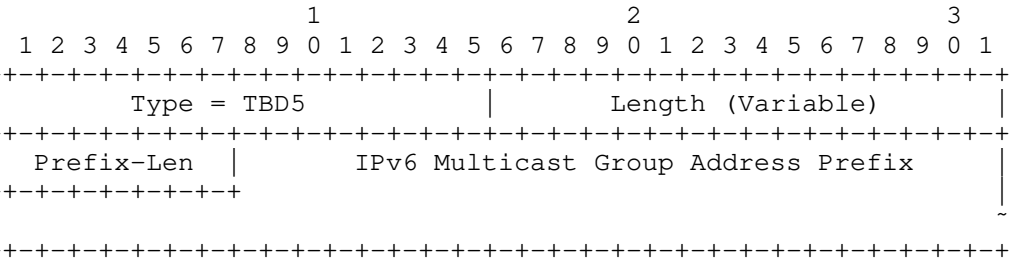


Figure 5: IPv6 Multicast Group Address Prefix sub-TLV

Type: TBD5 is to be assigned by IANA.

Length: Variable.

Prefix-Len: Indicates the length in bits of the IPv6 Multicast Group Address Prefix.

IPv6 Multicast Group Address Prefix: Indicates an IPv6 multicast group address prefix.

An IPv4 or IPv6 multicast source address prefix sub-TLV describes the source of the multicast traffic (or the packets). It is an IPv4 or IPv6 multicast source address prefix. The traffic (or the packets) with the IPv4 or IPv6 multicast group address from the source matching the prefix given in the IPv4 or IPv6 multicast source address prefix sub-TLV respectively will be transported by the BIER-TE path/tunnel.

The formats of an IPv4 and IPv6 multicast source address prefix sub-TLV are illustrated below.

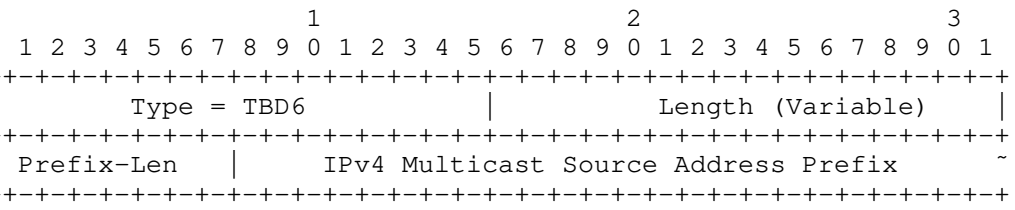


Figure 6: IPv4 Multicast Source Address Prefix sub-TLV

Type: TBD6 is to be assigned by IANA.

Length: Variable.

Prefix-Len: Indicates the length in bits of the IPv4 Multicast Source Address Prefix.

IPv4 Multicast Source Address Prefix: Indicates an IPv4 multicast source address prefix.

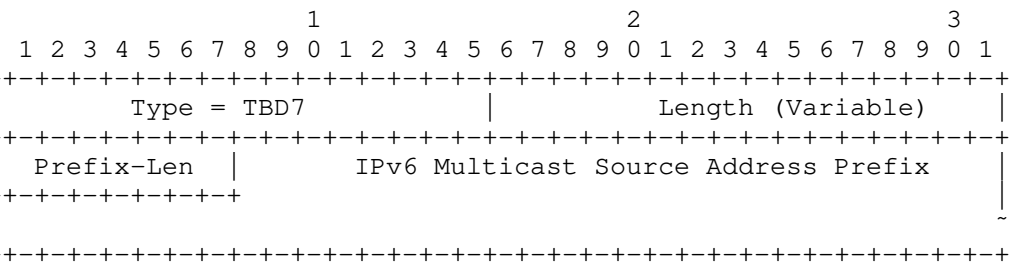


Figure 7: IPv6 Multicast Source Address Prefix sub-TLV

Type: TBD7 is to be assigned by IANA.

Length: Variable.

Prefix-Len: Indicates the length in bits of the IPv6 Multicast Source Address Prefix.

IPv6 Multicast Source Address Prefix: Indicates an IPv6 multicast source address prefix.

3.3. Ingress Node Object

To represent an ingress node, a new ingress node object is defined. The format of the new object for IPv4 (OT = 1) is as follows:

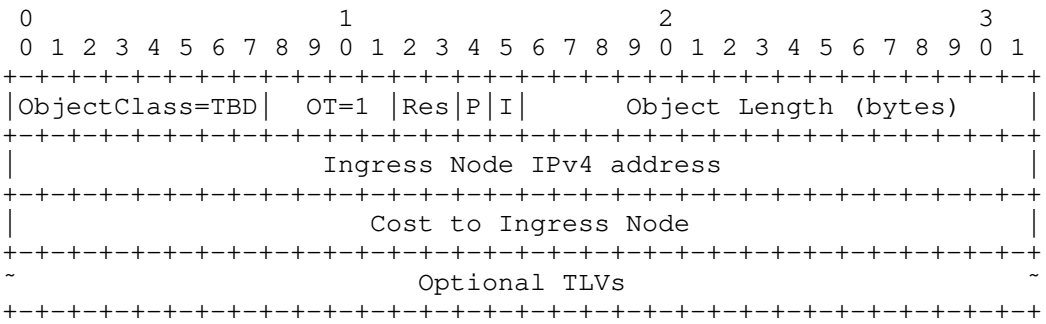


Figure 8: Ingress Node Object for IPv4

ObjectClass: TBD is to be assigned by IANA.

OT: 1 for IPv4.

Res, P, I and Object Length: Same as those defined in Common Object Header in [RFC5440].

Ingress Node IPv4 address: Indicates an IPv4 address of an ingress node.

Cost to Ingress Node: Indicates the cost from the multicast source to the ingress node.

No optional TLV is defined so far.

The format of the new object for IPv6 (OT = 2) is as follows:

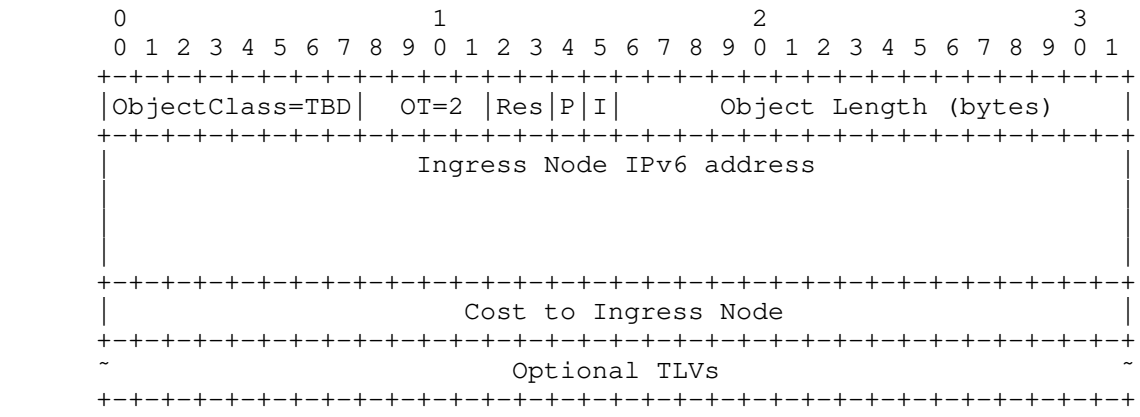


Figure 9: Ingress Node Object for IPv6

TBD, Res, P, I, Object Length, and Cost to Ingress Node:
Same as those defined in Ingress Node Object for IPv4.

OT: 2 for IPv6.

Ingress Node IPv6 address: Indicates an IPv6 address of an ingress node.

No optional TLV is defined so far.

3.4. Objective Functions

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

Objective Function Code: TBD8
Name: Minimum Bit Sets (MBS)
Description: Find a path represented by BitPositions that has the minimum number of bit sets.

Objective Function Code: TBD9
Name: Minimum Bits (MB)
Description: Find a path represented by BitPositions that has the minimum bit distance. The bit distance of BitPositions is the distance from the lowest bit to the highest bit in BitPositions.

3.5. BIER-TE Path Subobject

A BIER-TE path is represented by the BitPositions for the adjacencies through which the path traverses. A BitPosition is represented by a SI:BitString or a number.

A new subobject, called BIER-TE Path subobject (or BIER-TE-ERO subobject), is defined to contain the information about one or more BitPositions.

The format of a BIER-TE Path subobject in a ERO is shown in the figure below.

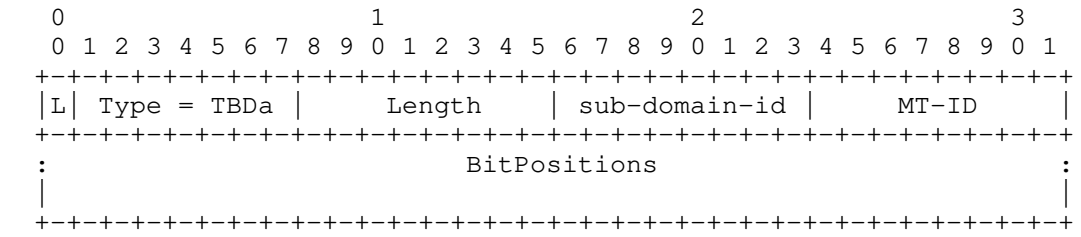


Figure 10: BIER-TE Path Subobject in ERO

L Flag (1 bit): It indicates whether the subobject represents a loose-hop in the path.

Type (7 bits): It is to be assigned by IANA. It identifies the BIER subobject type.

Length (8 bits): It contains the total length of the subobject in octets. The Length MUST be at least 4.

sub-domain-id: Unique value identifying the BIER sub-domain within the BIER domain.

MT-ID: Multi-Topology ID identifying the topology that is associated with the BIER sub-domain.

BitPositions: It MUST be at least one BitPosition.

For the subobject in a message received from a PCEP session, the format of the BitPositions in the subobject is determined by the values of SILen and BitStrLen in the PCE-BIER-TE-Path-Capability sub-TLV exchanged during the establishment of the session. When both SILen and BitStrLen are greater than zero, each of the BitPositions has two parts SI and BitString, where SI occupies SILen bits and

BitString occupies BitStrLen bits. When both SILen and BitStrLen are zeros, each of the BitPositions is a number of 16 bits.

For example, when SILen = 8 and BitStrLen = 1 (indicating BitString is of 64 bits), each BitPosition has a SI of 8 bits and a BitString of 64 bits. For simplicity, BitString of 8 bits is used below. The BitPositions for a BIER-TE path are sorted in descending order before they are put into a BIER-TE path subobject. For BIER-TE path {2', 4', 6', 16', 18', 2, 4}, when its BitPositions are sorted, it is {18', 16', 6', 4', 2', 4, 2}, which is {18'(8:00000010), 16'(7:10000000), 6'(6:00100000), 4'(6:00001000), 2'(6:00000010), 4 (0:00001000), 2 (0:00000010)}. The BitPositions with the same SI are stored in one BitString. For example, 6'(6:00100000), 4'(6:00001000) and 2'(6:00000010) are stored in (SI:BitString) = (6:00101010), where SI = 6. BIER-TE path {18', 16', 6', 4', 2', 4, 2} is encoded in the BIER-TE path subobject in the figure below. The path uses four BitStrings of 8 bits.

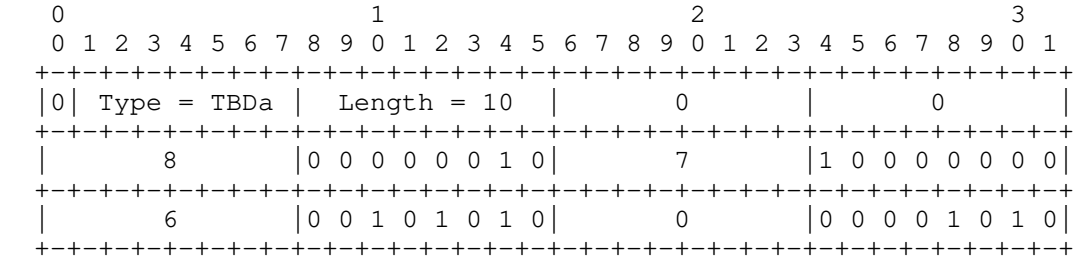


Figure 11: BIER-TE Path Subobject for a Path

3.6. BIER-TE Path Subobject in ERO

The ERO defined in [RFC5440] may contain a BIER-TE Path subobject for the BitPositions of a BIER-TE path. The BitPositions in the BIER-TE Path subobject for the BIER-TE path MUST be in descending order. When an ERO contains one or more BIER-TE Path subobjects for a BIER-TE path, the ERO MUST NOT include any other type of subobjects (i.e., it MUST include only BIER-TE Path subobjects). The first one is used and the others are ignored.

3.7. BIER-TE Path Subobject in RRO

A BIER-TE Path Subobject in a RRO (Record Route Object) has the same format as a BIER-TE Path subobject in a ERO except for L flag. The former does not have L flag. The format of a BIER-TE Path Subobject in a RRO is shown in the firgure below.

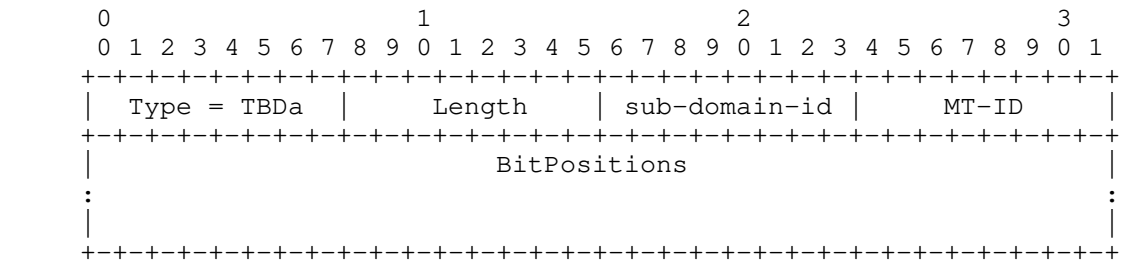


Figure 12: BIER-TE Path Subobject in RRO

A PCC may send a PCE a message such as a PCRpt message defined in [RFC8231]. The message contains a RRO with one BIER-TE Path subobject having the BitPositions for the actual BIER-TE path that is used to transport the traffic in the BIER-TE domain. The BitPositions in the BIER-TE Path subobject for the BIER-TE path MUST be in descending order.

4. Procedures

This section describes the procedures related to a BIER-TE path.

4.1. BIER-TE Path Creation

For PCC-Initiated BIER-TE path, a PCC MUST delegate the path by sending a path computation report (PCRpt) message with its demanded resources to a stateful PCE. Note the PCC MAY use the PCReq/PCRep before delegating.

Upon receiving the delegation via PCRpt message, the stateful PCE MUST compute a path based on the network resource availability stored in the TED.

The stateful PCE will send a PCUpd message for the BIER-TE path to the PCC. The stateful PCE MUST update its local LSP-DB and TED and would need to synchronize the information with other PCEs in the domain.

For PCE-Initiated BIER-TE path, the stateful PCE MUST compute a BIER-TE path per request from network management systems or applications automatically based on the network resource availability in the TED and send a PCInitiate message with the path information to the PCC. After receiving the PCInitiate message, the PCC creates the BIER-TE path.

For both PCC-Initiated and PCE-Initiated BIER-TE paths:

- The stateful PCE MUST update its local LSP-DB and TED with the paths.

- Upon receiving the PCUpd message or PCInitiate message for the path from the PCE with a found path, the PCC determines that it is a BIER-TE path by the PST = TBD1 for path setup using BIER-TE in the PATH-SETUP-TYPE TLV of the SRP object in the message.

4.2. BIER-TE Path Update

After a BIER-TE path is created in a BIER-TE domain, when some network events such as a node failure happen on the path (called old path) or a leaf/egress joins/leaves, the PCE computes a new BIER-TE path and replaces the old path with the new path. The new path satisfies the same constraints as the old path.

The PCE sends a PCUpd message to the PCC running on the ingress node. The message contains the information about the new BIER-TE path. After receiving the message, the PCC overwrites (or replaces) the BIER-TE path with the new BIER-TE path.

4.3. BIER-TE Path Deletion

For a BIER-TE path that has been created in a BIER-TE domain, after receiving a request for deleting the path from a user or application, the PCE MUST send a PCInitiate or PCUpd message to the PCC running on the ingress node of the path to remove the path.

5. The PCEP Messages

5.1. The PCRpt Message

The Path Computation State Report (PCRpt) message is a PCEP message sent by a PCC to a PCE to report the status of one or more LSPs, as per [RFC8281]. Each LSP State Report in a PCRpt message contains the actual LSP's path, bandwidth, operational and administrative status, etc. An LSP Status Report carried in a PCRpt message is also used in delegation or revocation of control of an LSP to/from a PCE.

In the case of a BIER-TE path, a PATH-SETUP-TYPE TLV with PST = TBD1 for path setup using BIER-TE MUST be carried in the SRP object in the PCRpt message. A BIER-TE path in the message is represented by a BIER-TE path subobject.

In addition, a PCRpt message is sent from the PCC running on an edge node to the PCE to report that the edge node as leaf/egress joins/ leaves to/from a multicast group and source.

5.2. The PCUpd Message

The Path Computation Update Request (PCUpd) message is a PCEP message sent by a PCE to a PCC to update LSP parameters on one or more LSPs, as per [RFC8281]. In the case of a BIER-TE path, a PATH-SETUP-TYPE TLV with PST = TBD1 for path setup using BIER-TE MUST be carried in the SRP object in the PCUpd message. Each BIER-TE path Update Request in a PCUpd message contains all parameters that a PCE wishes to be set for a given BIER-TE path. A BIER-TE path in the message is represented by a BIER-TE path subobject.

5.3. The PCInitiate Message

The LSP Initiate Request (PCInitiate) message is a PCEP message sent by a PCE to a PCC to trigger LSP instantiation or deletion, as per [RFC8281]. In the case of a BIER-TE path, a PATH-SETUP-TYPE TLV with PST = TBD1 for path setup using BIER-TE MUST be carried in the SRP object in the PCInitiate message. A BIER-TE path in the message is represented by a BIER-TE path subobject. The multicast packets to be transported by the BIER-TE path is specified by the Multicast Traffic TLV included in the SRP object.

5.4. The PCReq Message

The Path Computation Request (PCReq) message is a PCEP message sent by a PCC to a PCE to request a path computation [RFC5440], and it may contain the LSP object [RFC8231] to identify the LSP for which the path computation is requested. In the case of a BIER-TE path, a PATH-SETUP-TYPE TLV with PST = TBD1 for path setup using BIER-TE MUST be carried in the SRP object in the PCReq message.

In addition, a PCReq message is sent from the PCE (as a PCC) for the BIER-TE domain to another PCE for the domain that may contain the multicast source for a BIER-TE path in order to find an ingress node for the BIER-TE path.

5.5. The PCRep Message

The Path Computation Reply (PCRep) message is a PCEP message sent by a PCE to a PCC in reply to a path computation request [RFC5440], and it may contain the LSP object [RFC8231] to identify the LSP for which the path is computed. A PCRep message can contain either a set of computed paths if the request can be satisfied or a negative reply if not. A negative reply may indicate the reason why no path could be found. In the case of a BIER-TE path, a PATH-SETUP-TYPE TLV with PST = TBD1 for path setup using BIER-TE MUST be carried in the SRP object in the PCRep message. Each of the computed paths in the message is represented by a BIER-TE path subobject.

In addition, a PCRep message is sent from the PCE for the domain that may contain the multicast source for a BIER-TE path to the PCC (i.e., the PCE for the BIER-TE domain) in response to the request for finding an ingress node for the BIER-TE path. A PCRep message can contain either a set of ingress nodes represented by ingress node objects if the request can be satisfied or a negative reply if not.

6. IANA Considerations

6.1. PST for BIER-TE Path

IANA is requested to allocate a new code point within registry "PCEP Path Setup Types" under "Path Computation Element Protocol (PCEP) Numbers" as follows:

Value	Description	Reference
TBD1 (2)	Path is setup using BIER-TE	This document

6.2. PCE-BIER-TE-Path Capability sub-TLV

IANA is requested to allocate a new code point within registry "PATH-SETUP-TYPE-CAPABILITY Sub-TLV Type Indicators" under "Path Computation Element Protocol (PCEP) Numbers" as follows:

Value	Meaning	Reference
TBD2 (1)	PCE-BIER-TE-Path Capability	This document

6.3. SRP Object Flag Field

IANA is requested to allocate the following bits in the "SRP Object Flag Field" subregistry under the "Path Computation Element Protocol (PCEP) Numbers" registry:

Value	Description	Reference
27-29	Assistant Operations for Path	This document

6.4. Multicast Traffic TLV

IANA is requested to allocate the following value in the "PCEP TLV Type Indicators" subregistry under the "Path Computation Element Protocol (PCEP) Numbers" registry:

Value	Description	Reference
TBD3 (55)	Multicast Traffic	This document

IANA is requested to create and maintain a new sub-registry named "Multicast Traffic Sub-TLV Types" under the "Path Computation Element Protocol (PCEP) Numbers" registry. Initial values for the sub-registry are given below.

Type	Name	Reference
0	Reserved	This document
1	IPv4 Multicast Group Address Prefix	This document
2	IPv6 Multicast Group Address Prefix	This document
3	IPv4 Multicast Source Address Prefix	This document
4	IPv6 Multicast Source Address Prefix	This document
5-65535	Unassigned	This document

6.5. Ingress Node Object

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

Object-Class Value	Name	Object-Type	Reference
TBD (45)	INGRESS	0: Reserved	This document
		1: IPv4 Address	This document
		2: IPv6 Address	This document
		3-15:Unassigned	

6.6. OF Code Points

IANA is requested to allocate the following Objective Function Code Points in the "Objective Function" subregistry under the "Path Computation Element Protocol (PCEP) Numbers" registry:

Code Point	Name	Reference
TBD8 (18)	Minimum Bit Sets (MBS)	This document
TBD9 (19)	Minimum Bit Distance (MBD)	This document

6.7. PCEP BIER-TE Path Subobjects

This document defines a new subobject, called PCE BIER-TE Path (or BIER-TE-ERO) subobject, for PCEP ERO object. It also defines a new subobject, called PCE BIER-TE Path (or BIER-TE-RRO) subobject, for PCEP RRO object. The code points of the subobjects for the objects are maintained under ERO and RRO objects in the RSVP Parameters registry.

IANA is requested to allocate a code point under "Subobject type - 20 EXPLICIT_ROUTE - Type 1 Explicit Route" within registry "Resource Reservation Protocol (RSVP) Parameters" for PCEP BIER-TE path subobject as follows:

Value	Name	Reference
TBDa (63)	PCEP BIER-TE Path	This document

IANA is requested to allocate a code point under "Subobject type - 21 ROUTE_RECORD - Type 1 Explicit Route" within registry "Resource Reservation Protocol (RSVP) Parameters" for PCEP BIER-TE path subobject as follows:

Value	Name	Reference
TBDa (63)	PCEP BIER-TE Path	This document

7. Security Considerations

TBD

8. Acknowledgements

TBD

9. References

9.1. Normative References

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

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

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

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

[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>.

[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>.

9.2. Informative References

[I-D.chen-pce-bier]
Chen, R., Zhang, Z., Dhanaraj, S., and F. Qin, "PCEP Extensions for BIER-TE", draft-chen-pce-bier-08 (work in progress), November 2020.

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

Authors' Addresses

Huaimo Chen
Futurewei
Boston, MA
USA

Email: Huaimo.chen@futurewei.com

Mike McBride
Futurewei

Email: michael.mcbride@futurewei.com

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

Email: wangaj3@chinatelecom.cn

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

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

Yisong Liu
China Mobile

Email: liuyisong@chinamobile.com

Yanhe Fan
Casa Systems
USA

Email: yfan@casa-systems.com

Lei Liu
Fujitsu

USA

Email: liulei.kddi@gmail.com

Xufeng Liu
Volta Networks

McLean, VA
USA

Email: xufeng.liu.ietf@gmail.com

Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: October 17, 2021

H. Chen
M. McBride
Futurewei
A. Wang
China Telecom
G. Mishra
Verizon Inc.
Y. Liu
China Mobile
Y. Fan
Casa Systems
L. Liu
Fujitsu
X. Liu
Volta Networks
April 15, 2021

PCE for BIER-TE Path
draft-chen-pce-bier-te-path-01

Abstract

This document describes extensions to Path Computation Element (PCE) communication Protocol (PCEP) for supporting Bit Index Explicit Replication (BIER) Traffic Engineering (TE) paths.

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 October 17, 2021.

Table of Contents

1. Introduction

[I-D.ietf-bier-te-arch] introduces Bit Index Explicit Replication (BIER) Traffic/Tree Engineering (BIER-TE). It is an architecture for per-packet stateless explicit point to multipoint (P2MP) multicast path/tree and based on the BIER architecture defined in [RFC8279].

A Bit-Forwarding Ingress Router (BFIR) in a BIER-TE domain receives the information or instructions from a controller such as a stateful PCE about which multicast flows/packets are mapped to which P2MP paths. The multicast flows/packets are indicated by multicast and source addresses. The paths are represented by BitPositions or say BitStrings. After receiving the information or instructions, the ingress node/router encapsulates the multicast packets with the BitPositions for the corresponding P2MP paths, replicates and forwards the packets with the BitPositions along the P2MP paths.

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

To compute and initiate BIER-TE P2MP paths, the stateful PCE needs to be extended. For a BIER-TE P2MP path, some new state information will be stored and maintained, which includes the BitPositions, multicast group and multicast source for the path. The PCE gets the egresses of the path, the same multicast group and source from the egresses when each of the egresses reports to the PCE that it receives a multicast join with the multicast group and source. With this information, the PCE finds an ingress for the path, computes the path from the ingress to the egresses that has the optimal BitPositions and satisfies the constraints, and then initiates the BIER-TE path at the ingress of the path through sending the ingress the BitPositions of the path, multicast group and source in a PCEP message such as PCInitiate. After receiving the message, the ingress

creates a forwarding entry that imports the packets with the multicast group/address and source into the BIER-TE path (i.e., encapsulates the packets with a BIER-TE header having the BitPositions of the path), and then reports the status of the path to the PCE in a PCEP message such as PCRpt.

[I-D.chen-pce-bier] describes part of the solution for this, which is mainly the BIER-ERO subobject used for P2MP paths.

This document proposes a comprehensive solution for computing and establishing BIER-TE P2MP paths.

1.1. Terminologies

The following terminologies are used in this document.

PCE: Path Computation Element

PCEP: PCE communication Protocol

PCC: Path Computation Client

CE: Customer Edge

PE: Provider Edge

BIER: Bit Index Explicit Replication.

BIER-TE: BIER Traffic/Tree Engineering.

BFR: Bit-Forwarding Router.

BFIR: Bit-Forwarding Ingress Router.

BFER: Bit-Forwarding Egress Router.

BFR-id: BFR Identifier. It is a number in the range [1,65535].

BFR-NBR: BFR Neighbor.

BFR-prefix: An IP address (either IPv4 or IPv6) of a BFR.

BIRT: Bit Index Routing Table. It is a table that maps from the BFR-id (in a particular sub-domain) of a BFER to the BFR-prefix of that BFER, and to the BFR-NBR on the path to that BFER.

BIFT: Bit Index Forwarding Table.

LSP-DB: Label Switching Path DataBase.

TED: Traffic/Tree Engineering DataBase.

2. Overview of PCE for BIER-TE

This section briefly describes PCE for BIER-TE and illustrates some details through a simple example BIER-TE topology.

2.1. Example BIER-TE Topology with PCE

An example BIER-TE topology for a BIER-TE domain with a PCE is shown in Figure 1. There are 8 nodes/BFRs A, B, C, D, E, F, G and H in the domain. Nodes/BFRs A, H, E, F and D are BFIRs (i.e., ingress nodes) or BFERs (i.e., egress nodes). There is a connection (i.e., PCE session) between the PCE and the PCC running on each of the possible ingress and egress nodes in the domain. Note that some of connections and the PCC on each node are not shown in the figure.

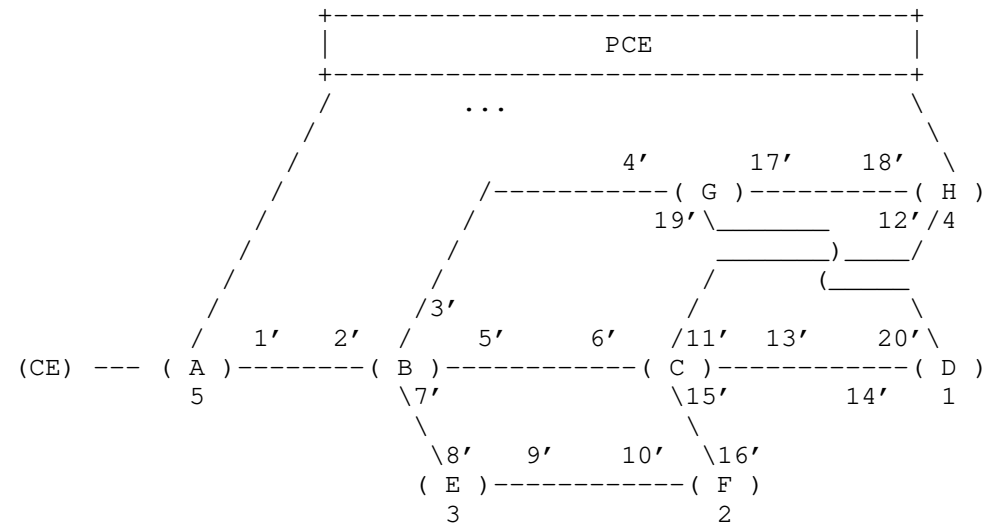


Figure 1: Example BIER-TE Topology with PCE

Nodes/BFRs D, F, E, H and A are BFERs (or BFIRs) and have local decap adjacency BitPositions 1, 2, 3, 4, and 5 respectively. For simplicity, these BPs are represented by (SI:BitString), where SI = 0 and BitString is of 8 bits. BPs 1, 2, 3, 4, and 5 are represented by 1 (0:00000001), 2 (0:00000010), 3 (0:00000100), 4 (0:00001000) and 5 (0:00010000) respectively.

The BitPositions for the forward connected adjacencies are represented by i', where i is from 1 to 20. In one option, they are encoded as (n+i), where n is a power of 2 such as 32768. For simplicity, these BitPositions are represented by (SI:BitString), where SI = (6 + (i-1)/8) and BitString is of 8 bits. BitPositions i' (i from 1 to 20) are represented by 1'(6:00000001), 2'(6:00000010), 3'(6:00000100), 4'(6:00001000), 5'(6:00010000), 6'(6:00100000), 7'(6:01000000), 8'(6:10000000), 9'(7:00000001), 10'(7:00000010), . . . , 16'(7:10000000), 17'(8:00000001), 18'(8:00000010), . . . , 20'(8:00001000).

For a link between two nodes X and Y, there are two BitPositions for two forward connected adjacencies. These two forward connected adjacency BitPositions are assigned on nodes X and Y respectively. The BitPosition assigned on X is the forward connected adjacency of Y. The BitPosition assigned on Y is the forward connected adjacency of X.

For example, for the link between nodes B and C in the figure, two forward connected adjacency BitPositions 5' and 6' are assigned to two ends of the link. BitPosition 5' is assigned on node B to B's end of the link. It is the forward connected adjacency of node C. BitPosition 6' is assigned on node C to C's end of the link. It is the forward connected adjacency of node B.

2.2. A Brief Flow of PCEP Messages for a BIER-TE Path

For a BIER-TE Path to transport the packets with a given multicast group/address and source in a BIER-TE domain, a sequence of PCEP messages are exchanged between the PCE for the domain and the PCEs for the domains containing the source, and between the PCE for the domain and the PCCs running on the BFERs/BFIRs of the domain.

Suppose that each of nodes H, D and F receives a multicast join with a same multicast group/address and source, which are MGa and MSa respectively. For simplicity, assume that the multicast source MSa is in the left domain containing the CE in Figure 1. The following is a brief flow of PCEP messages for computing and creating a BIER-TE Path to transport the packets to H, D and F.

At first, the PCC running on each of nodes H, D and F sends the PCE a PCEP message such as PCRpt. The message contains the multicast group and source (i.e., MGa and MSa), which reports to the PCE that the node receives a multicast join with MGa and MSa. Note that a PCEP message is sent to the PCE from the PCC on a node to report that the node leaves when the node receives a multicast leave with MGa and MSa.

After receiving the PCEP messages from nodes H, D and F reporting multicast join with MGa and MSa, the PCE for the domain containing these nodes determines that nodes H, D and F are the egress nodes of a BIER-TE path since they have the same multicast group and source.

Second, the PCE for the domain sends a PCEP message such as PCReq to each of the PCEs for the domains that may contain the multicast source. This message requests the PCE (that may contain the source) to find an ingress node for the BIER-TE path having egress nodes H, D and F. The message contains the multicast group and source (i.e., MGa and MSa). For example, the PCE for the BIER-TE domain sends the PCEP message to the PCE (called PCE-L) for the left domain containing CE (note that this PCE is not shown in the figure).

After receiving the PCEP message requesting to find an ingress node, the PCE (e.g., PCE-L) for the domain containing the multicast source computes the ingress node that is reachable from the source with minimum cost (e.g., ingress node A). The PCE for the domain without the source can not find any ingress node.

Third, the PCE for the domain with the source sends the PCE for the BIER-TE domain a PCEP message such as PCRep with the ingress node. The PCE for the domain without the source sends the PCE for the BIER-TE domain a PCEP message such as PCRep with NO INGRESS FOUND.

After receiving the PCEP message with the ingress node, the PCE for the BIER-TE domain computes a P2MP path from the ingress node (e.g., A) to the egress nodes (e.g., H, D and F). The path has the optimal BitPositions and satisfies the constraints. The optimal BitPositions means the BitPositions for the path has the minimum number of bit sets and the minimum bit distance.

Fourth, the PCE for the BIER-TE domain sends a PCEP message such as PCInitiate to the PCC on the ingress node (e.g., A) for the ingress to create a BIER-TE path to transport the packets for the given multicast group and source. The message contains the BitPositions for the path, the multicast group and source.

After receiving the PCEP message with the path, the PCC on the ingress (e.g., A) creates the BIER-TE path, i.e., a forwarding entry that imports the packets with the multicast group/address and source into the BIER-TE path (i.e., encapsulates the packets with a BIER-TE header having the BitPositions of the path).

And then the PCC on the ingress sends the PCE a PCEP message such as PCRpt reporting the status of the path to the PCE.

After receiving the PCEP message with the status of the path, the PCE for the domain updates the information about the path accordingly.

2.3. Procedures on Ingress

This section introduces the procedures for the ingress node of a P2MP path to get the BitPositions representing the explicit P2MP path from the ingress node to its egress nodes from the PCE.

Suppose that node A in Figure 1 wants to have an explicit P2MP path from ingress node A to egress nodes H and F. The path satisfies a set of constraints. In one case, the PCC running on ingress node A sends a request for the path to the PCE. The request contains the set of constraints, objective functions, the ingress node and the egress nodes. After receiving the request, the PCE computes an explicit P2MP path, which satisfies the constraints and is from the given ingress node to the egress nodes. While computing the path, the PCE will optimize the BitPositions of the path. That is that, for a given length of BitString, the path computed uses the minimum number of BitStrings (i.e., bit sets) and satisfies the constraints. The length is given by the value in BitStrLen field in the PCE-BIER-TE-Path-Capability sub-TLV. The PCE sends a reply with the path to the PCC. The reply contains the BitPositions representing the explicit P2MP path.

For example, assume that the explicit P2MP path computed by the PCE traverses the link/adjacency from A to B (indicated by BP 2’), the link/adjacency from B to G (indicated by BP 4’) and the link/adjacency from B to C (indicated by BP 6’), the link/adjacency from G to H (indicated by BP 18’), and the link/adjacency from C to F (indicated by BP 16’). This path is represented by {2’, 4’, 6’, 16’, 18’, 2, 4}, where BitPositions 2 and 4 indicate egress nodes F and H respectively. The reply sent to the PCC on node A by the PCE contains the path represented by {2’, 4’, 6’, 16’, 18’, 2, 4}.

In another case, a request for a P2MP path is from a user or application. After receiving the request, the PCE finds an ingress node if no ingress is given, and computes an explicit P2MP path from the ingress node to the egress nodes and sends the path to the PCC running on the ingress node.

After receiving the P2MP path, for any packet from CE to be transported by the path, such as the packet with the multicast address, the ingress node encapsulates the packet with the BitPositions representing the path and forwards the packet according to its BIFT.

For example, when ingress node A receives the path represented by BitPositions {2', 4', 6', 16', 18', 2, 4}, it encapsulates every packet from CE with the multicast address with the BitPositions and then forwards the packet along the P2MP path according to its BIFT.

A forwards the packet to B according to the forwarding entry for BP 2' in its BIFT.

After receiving the packet from A, B forwards the packet to G and C according to the forwarding entries for BPs 4' and 6' in B's BIFT respectively. The packet received by G has path {16', 18', 2, 4}. The packet received by C has path {16', 18', 2, 4}.

After receiving the packet from B, G sends the packet to H according to the forwarding entry for BP 18' in G's BIFT.

After receiving the packet from B, C sends the packet to F according to the forwarding entry for BP 16' in C's BIFT.

Egress node H of the P2MP path receives the packet with BitPosition 4. It decapsulates the packet and pass the payload of the packet to the packet's NextProto.

Egress node F of the P2MP path receives the packet with BitPosition 2. It decapsulates the packet and pass the payload of the packet to the packet's NextProto.

3. Extensions to PCEP

This section describes extensions to PCEP.

3.1. BIER-TE Path Capability

During a PCEP session establishment, PCEP Speakers (PCE or PCC) indicate their ability to support BIER-TE paths. The OPEN object in the Open message contains the PATH-SETUP-TYPE-CAPABILITY TLV, which is defined in [RFC8408]. The TLV contains a list of Path Setup Types (PSTs) and optional sub-TLVs associated with the PSTs. The sub-TLVs convey the parameters that are associated with the PSTs supported by a PCEP speaker.

This document defines a new PST value:

* PST = TBD1: Path is setup using BIER-TE.

A new sub-TLV associated with this new PST is defined, which is called PCE-BIER-TE-Path-Capability sub-TLV. The format of this new sub-TLV is illustrated in the figure below.

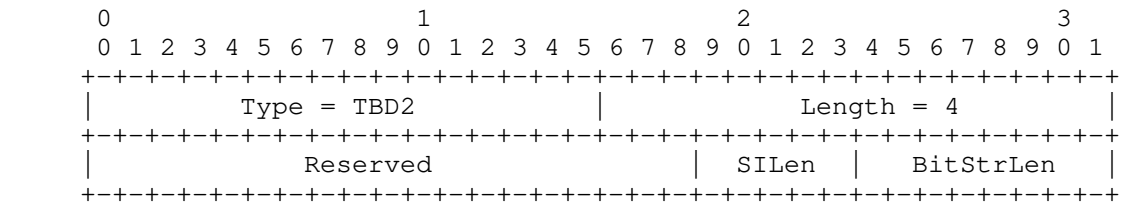


Figure 2: PCE-BIER-TE-Path-Capability sub-TLV

Type - 16 bits: TBD2 is to be assigned by IANA.

Length - 16 bits: 4 is the total length in bytes of the remainder of the TLV, excluding the Type and Length fields.

SILen (SI Length) - 5 bits: The length in bits of the SI field.

BitStrLen (Bit String Length) - 8 bits: The length in bits of the BitString field according to [RFC8296]. If k is the length of the BitString, the value of BitStrLen is log2(k)-5. For example, BitStrLen = 1 indicates k = 64, BitStrLen = 7 indicates k = 4096.

Reserved - 19 bits: MUST be set to zero by the sender and MUST be ignored by the receiver.

A PCEP speaker supporting BIER-TE paths includes the new PST and sub-TLV in the PATH-SETUP-TYPE-CAPABILITY TLV.

3.2. Extensions to SRP

For a PCEP message, when it is used for a BIER-TE path, the SRP (Stateful PCE Request Parameters) object in the message MUST include the PATH-SETUP-TYPE TLV defined in [RFC8408]. The TLV MUST contain the PST = TBD1 for path setup using BIER-TE.

Three contiguous bits in SRP Object Flag Field are defined to indicate one of the assistant operations for a BIER-TE path. This three bits field is called AOP (Assistant Operations). In addition, the Multicast Flow Specification TLV defined in [I-D.ietf-pce-pcep-flowspec] is re-used in the SRP object for indicating Multicast Traffic.

3.2.1. SRP Object Flag Field

The three bits for AOP are bits 27 to 29 (the exact bits to be assigned by IANA) in the SRP Object Flag Field. The values of AOP are defined as follows:

```
AOP Value    Meaning (Assistant Operation)
0x001 (J):  Join with Multicast Group and Source
0x010 (L):  Leave from Multicast Group and Source
0x011 (I):  Ingress node computation
```

The value of AOP indicates one of the three operations above. When any of the other values is received, an error MUST be reported.

When the PCC running on an edge node of a BIER-TE domain sends the PCE for the domain a PCEP message such as PCRpt to report that the edge node receives a multicast join, the message MUST include a SRP object with AOP == 0x001 (J).

When the PCC running on an edge node of a BIER-TE domain sends the PCE for the domain a PCEP message such as PCRpt to report that the edge node receives a multicast leave, the message MUST include a SRP object with AOP == 0x010 (L).

When the PCE for the domain sends a PCEP message such as PCReq to another PCE for requesting to find an ingress node for a BIER-TE path, the message MUST include a SRP object with AOP == 0x011 (I).

3.2.2. Reuse of Multicast Flow Specification TLV

For a PCE-Initiated BIER-TE path, when a PCE sends a PCC a message such as PCInitiate message to create a BIER-TE path in a BIER-TE domain, the message MUST contain a Multicast Flow Specification TLV in the SRP object. The TLV indicates the multicast traffic that the BIER-TE path will carry.

When the PCC running on an edge node of a BIER-TE domain sends the PCE for the domain a PCEP message to report that the edge node receives a multicast join or leave with a multicast group/address and source, the message MUST contain a Multicast Flow Specification TLV in the SRP object. The TLV indicates the multicast group by the multicast group adress and/or multicast source address.

When the PCE for a BIER-TE domain sends another PCE a PCEP message to request for finding an ingress node of a BIER-TE path, the message MUST contain a Multicast Flow Specification TLV in the SRP object. The TLV indicates the multicast source.

3.3. Ingress Node Object

To represent an ingress node, a new ingress node object is defined. The format of the new object for IPv4 (OT = 1) is as follows:

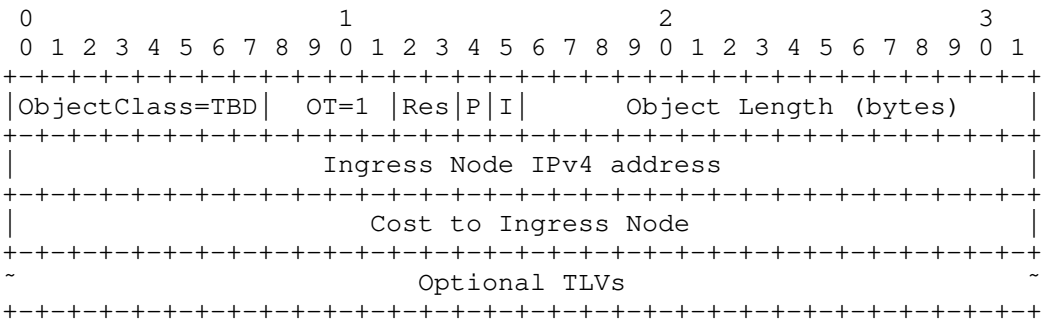


Figure 3: Ingress Node Object for IPv4

ObjectClass: TBD is to be assigned by IANA.

OT: 1 for IPv4.

Res, P, I and Object Length: Same as those defined in Common Object Header in [RFC5440].

Ingress Node IPv4 address: Indicates an IPv4 address of an ingress node.

Cost to Ingress Node: Indicates the cost from the multicast source to the ingress node.

No optional TLV is defined so far.

The format of the new object for IPv6 (OT = 2) is as follows:

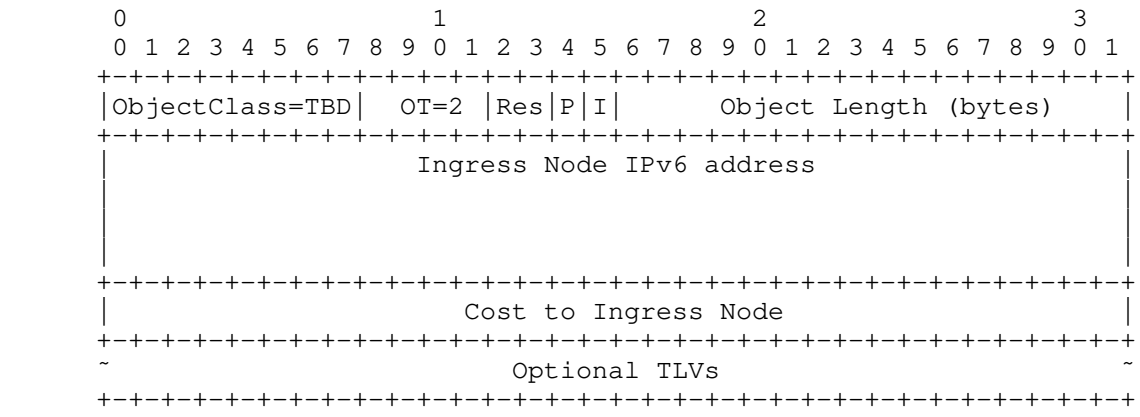


Figure 4: Ingress Node Object for IPv6

TBD, Res, P, I, Object Length, and Cost to Ingress Node:
Same as those defined in Ingress Node Object for IPv4.

OT: 2 for IPv6.

Ingress Node IPv6 address: Indicates an IPv6 address of an ingress node.

No optional TLV is defined so far.

3.4. Objective Functions

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

Objective Function Code: TBD8
Name: Minimum Bit Sets (MBS)
Description: Find a path represented by BitPositions that has the minimum number of bit sets.

Objective Function Code: TBD9
Name: Minimum Bits (MB)
Description: Find a path represented by BitPositions that has the minimum bit distance. The bit distance of BitPositions is the distance from the lowest bit to the highest bit in BitPositions.

3.5. BIER-TE Path Subobject

A BIER-TE path is represented by the BitPositions for the adjacencies through which the path traverses. A BitPosition is represented by a SI:BitString or a number.

A new subobject, called BIER-TE Path subobject (or BIER-TE-ERO subobject), is defined to contain the information about one or more BitPositions.

The format of a BIER-TE Path subobject in a ERO is shown in the figure below.

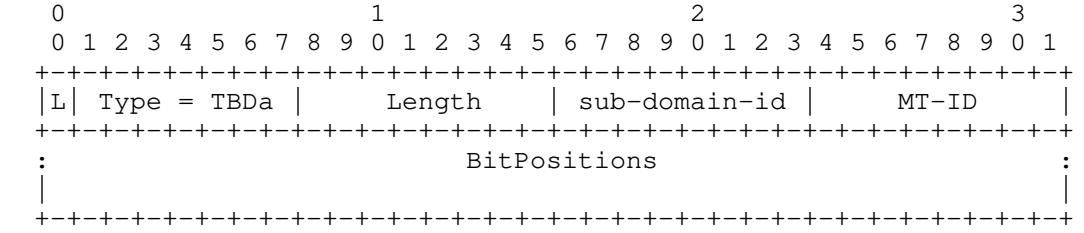


Figure 5: BIER-TE Path Subobject in ERO

L Flag (1 bit): It indicates whether the subobject represents a loose-hop in the path.

Type (7 bits): It is to be assigned by IANA. It identifies the BIER subobject type.

Length (8 bits): It contains the total length of the subobject in octets. The Length MUST be at least 4.

sub-domain-id: Unique value identifying the BIER sub-domain within the BIER domain.

MT-ID: Multi-Topology ID identifying the topology that is associated with the BIER sub-domain.

BitPositions: It MUST be at least one BitPosition.

For the subobject in a message received from a PCEP session, the format of the BitPositions in the subobject is determined by the values of SILen and BitStrLen in the PCE-BIER-TE-Path-Capability sub-TLV exchanged during the establishment of the session. When both SILen and BitStrLen are greater than zero, each of the BitPositions has two parts SI and BitString, where SI occupies SILen bits and

BitString occupies BitStrLen bits. When both SILen and BitStrLen are zeros, each of the BitPositions is a number of 16 bits.

For example, when SILen = 8 and BitStrLen = 1 (indicating BitString is of 64 bits), each BitPosition has a SI of 8 bits and a BitString of 64 bits. For simplicity, BitString of 8 bits is used below. The BitPositions for a BIER-TE path are sorted in descending order before they are put into a BIER-TE path subobject. For BIER-TE path {2', 4', 6', 16', 18', 2, 4}, when its BitPositions are sorted, it is {18', 16', 6', 4', 2', 4, 2}, which is {18'(8:00000010), 16'(7:10000000), 6'(6:00100000), 4'(6:00001000), 2'(6:00000010), 4 (0:00001000), 2 (0:00000010)}. The BitPositions with the same SI are stored in one BitString. For example, 6'(6:00100000), 4'(6:00001000) and 2'(6:00000010) are stored in (SI:BitString) = (6:00101010), where SI = 6. BIER-TE path {18', 16', 6', 4', 2', 4, 2} is encoded in the BIER-TE path subobject in the figure below. The path uses four BitStrings of 8 bits.

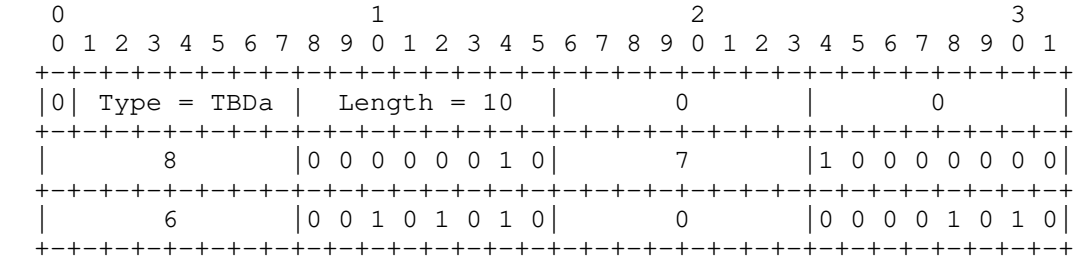


Figure 6: BIER-TE Path Subobject for a Path

3.6. BIER-TE Path Subobject in ERO

The ERO defined in [RFC5440] may contain a BIER-TE Path subobject for the BitPositions of a BIER-TE path. The BitPositions in the BIER-TE Path subobject for the BIER-TE path MUST be in descending order. When an ERO contains one or more BIER-TE Path subobjects for a BIER-TE path, the ERO MUST NOT include any other type of subobjects (i.e., it MUST include only BIER-TE Path subobjects). The first one is used and the others are ignored.

3.7. BIER-TE Path Subobject in RRO

A BIER-TE Path Subobject in a RRO (Record Route Object) has the same format as a BIER-TE Path subobject in a ERO except for L flag. The former does not have L flag. The format of a BIER-TE Path Subobject in a RRO is shown in the firgure below.

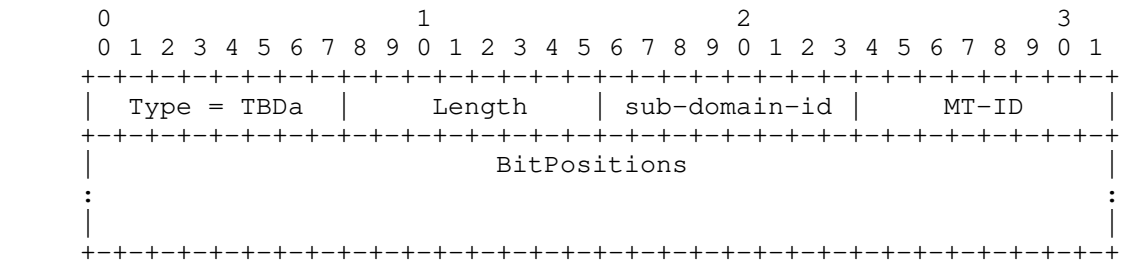


Figure 7: BIER-TE Path Subobject in RRO

A PCC may send a PCE a message such as a PCRpt message defined in [RFC8231]. The message contains a RRO with one BIER-TE Path subobject having the BitPositions for the actual BIER-TE path that is used to transport the traffic in the BIER-TE domain. The BitPositions in the BIER-TE Path subobject for the BIER-TE path MUST be in descending order.

4. Procedures

This section describes the procedures related to a BIER-TE path.

4.1. BIER-TE Path Creation

For PCC-Initiated BIER-TE path, a PCC MUST delegate the path by sending a path computation report (PCRpt) message with its demanded resources to a stateful PCE. Note the PCC MAY use the PCReq/PCRep before delegating.

Upon receiving the delegation via PCRpt message, the stateful PCE MUST compute a path based on the network resource availability stored in the TED.

The stateful PCE will send a PCUpd message for the BIER-TE path to the PCC. The stateful PCE MUST update its local LSP-DB and TED and would need to synchronize the information with other PCEs in the domain.

For PCE-Initiated BIER-TE path, the stateful PCE MUST compute a BIER-TE path per request from network management systems or applications automatically based on the network resource availability in the TED and send a PCInitiate message with the path information to the PCC. After receiving the PCInitiate message, the PCC creates the BIER-TE path.

For both PCC-Initiated and PCE-Initiated BIER-TE paths:

- The stateful PCE MUST update its local LSP-DB and TED with the paths.

- Upon receiving the PCUpd message or PCInitiate message for the path from the PCE with a found path, the PCC determines that it is a BIER-TE path by the PST = TBD1 for path setup using BIER-TE in the PATH-SETUP-TYPE TLV of the SRP object in the message.

4.2. BIER-TE Path Update

After a BIER-TE path is created in a BIER-TE domain, when some network events such as a node failure happen on the path (called old path) or a leaf/egress joins/leaves, the PCE computes a new BIER-TE path and replaces the old path with the new path. The new path satisfies the same constraints as the old path.

The PCE sends a PCUpd message to the PCC running on the ingress node. The message contains the information about the new BIER-TE path. After receiving the message, the PCC overwrites (or replaces) the BIER-TE path with the new BIER-TE path.

4.3. BIER-TE Path Deletion

For a BIER-TE path that has been created in a BIER-TE domain, after receiving a request for deleting the path from a user or application, the PCE MUST send a PCInitiate or PCUpd message to the PCC running on the ingress node of the path to remove the path.

5. The PCEP Messages

5.1. The PCRpt Message

The Path Computation State Report (PCRpt) message is a PCEP message sent by a PCC to a PCE to report the status of one or more LSPs, as per [RFC8281]. Each LSP State Report in a PCRpt message contains the actual LSP's path, bandwidth, operational and administrative status, etc. An LSP Status Report carried in a PCRpt message is also used in delegation or revocation of control of an LSP to/from a PCE.

In the case of a BIER-TE path, a PATH-SETUP-TYPE TLV with PST = TBD1 for path setup using BIER-TE MUST be carried in the SRP object in the PCRpt message. A BIER-TE path in the message is represented by a BIER-TE path subobject.

In addition, a PCRpt message is sent from the PCC running on an edge node to the PCE to report that the edge node as leaf/egress joins/ leaves to/from a multicast group and source.

5.2. The PCUpd Message

The Path Computation Update Request (PCUpd) message is a PCEP message sent by a PCE to a PCC to update LSP parameters on one or more LSPs, as per [RFC8281]. In the case of a BIER-TE path, a PATH-SETUP-TYPE TLV with PST = TBD1 for path setup using BIER-TE MUST be carried in the SRP object in the PCUpd message. Each BIER-TE path Update Request in a PCUpd message contains all parameters that a PCE wishes to be set for a given BIER-TE path. A BIER-TE path in the message is represented by a BIER-TE path subobject.

5.3. The PCInitiate Message

The LSP Initiate Request (PCInitiate) message is a PCEP message sent by a PCE to a PCC to trigger LSP instantiation or deletion, as per [RFC8281]. In the case of a BIER-TE path, a PATH-SETUP-TYPE TLV with PST = TBD1 for path setup using BIER-TE MUST be carried in the SRP object in the PCInitiate message. A BIER-TE path in the message is represented by a BIER-TE path subobject. The multicast packets to be transported by the BIER-TE path is specified by the Multicast Flow Specification TLV included in the SRP object.

5.4. The PCReq Message

The Path Computation Request (PCReq) message is a PCEP message sent by a PCC to a PCE to request a path computation [RFC5440], and it may contain the LSP object [RFC8231] to identify the LSP for which the path computation is requested. In the case of a BIER-TE path, a PATH-SETUP-TYPE TLV with PST = TBD1 for path setup using BIER-TE MUST be carried in the SRP object in the PCReq message.

In addition, a PCReq message is sent from the PCE (as a PCC) for the BIER-TE domain to another PCE for the domain that may contain the multicast source for a BIER-TE path in order to find an ingress node for the BIER-TE path.

5.5. The PCRep Message

The Path Computation Reply (PCRep) message is a PCEP message sent by a PCE to a PCC in reply to a path computation request [RFC5440], and it may contain the LSP object [RFC8231] to identify the LSP for which the path is computed. A PCRep message can contain either a set of computed paths if the request can be satisfied or a negative reply if not. A negative reply may indicate the reason why no path could be found. In the case of a BIER-TE path, a PATH-SETUP-TYPE TLV with PST = TBD1 for path setup using BIER-TE MUST be carried in the SRP object in the PCRep message. Each of the computed paths in the message is represented by a BIER-TE path subobject.

In addition, a PCRep message is sent from the PCE for the domain that may contain the multicast source for a BIER-TE path to the PCC (i.e., the PCE for the BIER-TE domain) in response to the request for finding an ingress node for the BIER-TE path. A PCRep message can contain either a set of ingress nodes represented by ingress node objects if the request can be satisfied or a negative reply if not.

6. IANA Considerations

6.1. PST for BIER-TE Path

IANA is requested to allocate a new code point within registry "PCEP Path Setup Types" under "Path Computation Element Protocol (PCEP) Numbers" as follows:

Value	Description	Reference
TBD1 (2)	Path is setup using BIER-TE	This document

6.2. PCE-BIER-TE-Path Capability sub-TLV

IANA is requested to allocate a new code point within registry "PATH-SETUP-TYPE-CAPABILITY Sub-TLV Type Indicators" under "Path Computation Element Protocol (PCEP) Numbers" as follows:

Value	Meaning	Reference
TBD2 (1)	PCE-BIER-TE-Path Capability	This document

6.3. SRP Object Flag Field

IANA is requested to allocate the following bits in the "SRP Object Flag Field" subregistry under the "Path Computation Element Protocol (PCEP) Numbers" registry:

Value	Description	Reference
27-29	Assistant Operations for Path	This document

6.4. Ingress Node Object

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

Object-Class Value	Name	Object-Type	Reference
TBD (45)	INGRESS	0: Reserved	This document
		1: IPv4 Address	This document
		2: IPv6 Address	This document
		3-15:Unassigned	

6.5. OF Code Points

IANA is requested to allocate the following Objective Function Code Points in the "Objective Function" subregistry under the "Path Computation Element Protocol (PCEP) Numbers" registry:

Code Point	Name	Reference
TBD8 (18)	Minimum Bit Sets (MBS)	This document
TBD9 (19)	Minimum Bit Distance (MBD)	This document

6.6. PCEP BIER-TE Path Subobjects

This document defines a new subobject, called PCE BIER-TE Path (or BIER-TE-ERO) subobject, for PCEP ERO object. It also defines a new subobject, called PCE BIER-TE Path (or BIER-TE-RRO) subobject, for PCEP RRO object. The code points of the subobjects for the objects are maintained under ERO and RRO objects in the RSVP Parameters registry.

IANA is requested to allocate a code point under "Subobject type - 20 EXPLICIT_ROUTE - Type 1 Explicit Route" within registry "Resource Reservation Protocol (RSVP) Parameters" for PCEP BIER-TE path subobject as follows:

Value	Name	Reference
TBDa (63)	PCEP BIER-TE Path	This document

IANA is requested to allocate a code point under "Subobject type - 21 ROUTE_RECORD - Type 1 Explicit Route" within registry "Resource Reservation Protocol (RSVP) Parameters" for PCEP BIER-TE path subobject as follows:

Value	Name	Reference
TBDa (63)	PCEP BIER-TE Path	This document

7. Security Considerations

TBD

8. Acknowledgements

The authors would like to thank Dhruv Dhody, and others for their comments to this work.

9. References

9.1. Normative References

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

[I-D.ietf-pce-pcep-flowspec]
Dhody, D., Farrel, A., and Z. Li, "PCEP Extension for Flow Specification", draft-ietf-pce-pcep-flowspec-12 (work in progress), October 2020.

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

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

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

[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>.

[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>.

9.2. Informative References

[I-D.chen-pce-bier]
Chen, R., Zhang, Z., Dhanaraj, S., and F. Qin, "PCEP Extensions for BIER-TE", draft-chen-pce-bier-08 (work in progress), November 2020.

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

Authors' Addresses

Huaimo Chen
Futurewei
Boston, MA
USA

Email: Huaimo.chen@futurewei.com

Mike McBride
Futurewei

Email: michael.mcbride@futurewei.com

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

Email: wangaj3@chinatelecom.cn

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

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

Yisong Liu
China Mobile

Email: liuyisong@chinamobile.com

Yanhe Fan
Casa Systems
USA

Email: yfan@casa-systems.com

Lei Liu
Fujitsu

USA

Email: liulei.kddi@gmail.com

Xufeng Liu
Volta Networks

McLean, VA
USA

Email: xufeng.liu.ietf@gmail.com

PCE
Internet-Draft
Intended status: Standards Track
Expires: August 13, 2021

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

February 9, 2021

Path Computation Element Communication Protocol (PCEP) Extensions to Enable IFIT
draft-chen-pce-pcep-ifit-02

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 13, 2021.

Table of Contents

1. Introduction

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

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

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

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

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

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

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

The Use Case of Segment Routing (SR) is 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). In this way SR policy native IFIT can facilitate the closed loop control and enable the automation of SR service.

Segment Routing (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.

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

2. PCEP Extensions for IFIT Attributes

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

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

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

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

2.1. IFIT for SR Policies

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

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

3. IFIT capability advertisement TLV

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

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

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

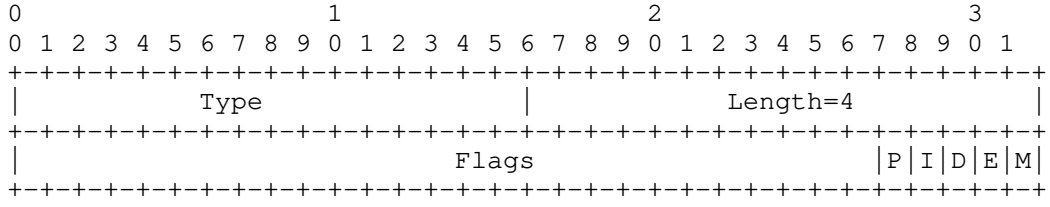


Fig. 1 IFIT-CAPABILITY TLV Format

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:

- 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.

- 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:

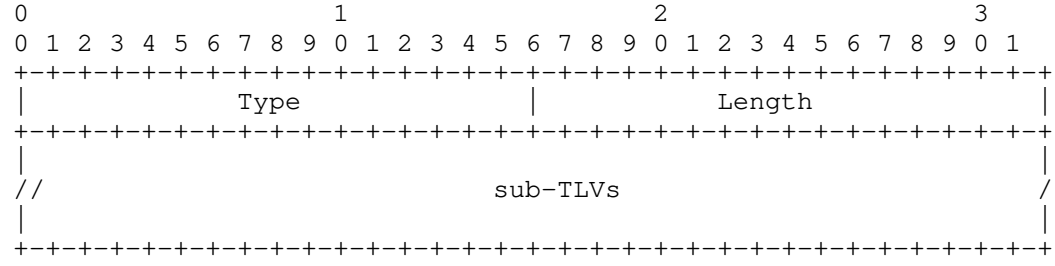


Fig. 2 IFIT-ATTRIBUTES TLV Format

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:

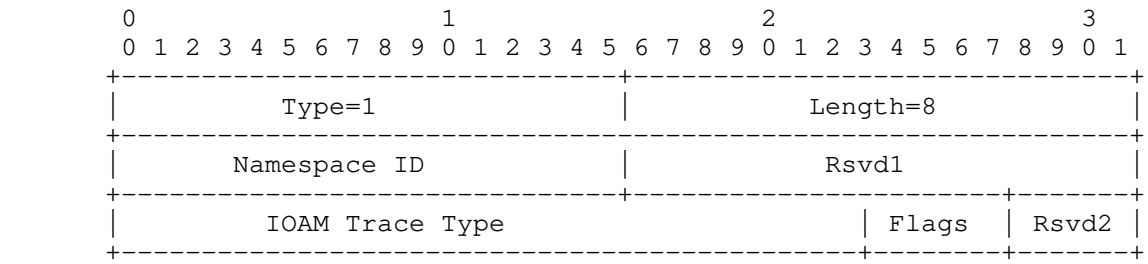


Fig. 4 IOAM Pre-allocated Trace Option Sub-TLV

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:

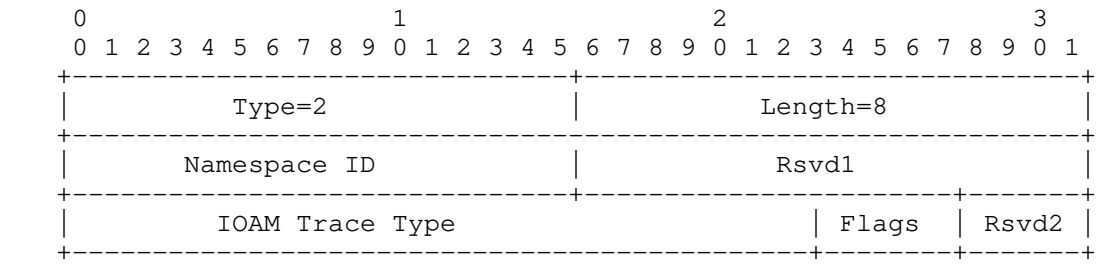


Fig. 5 IOAM Incremental Trace Option Sub-TLV

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:

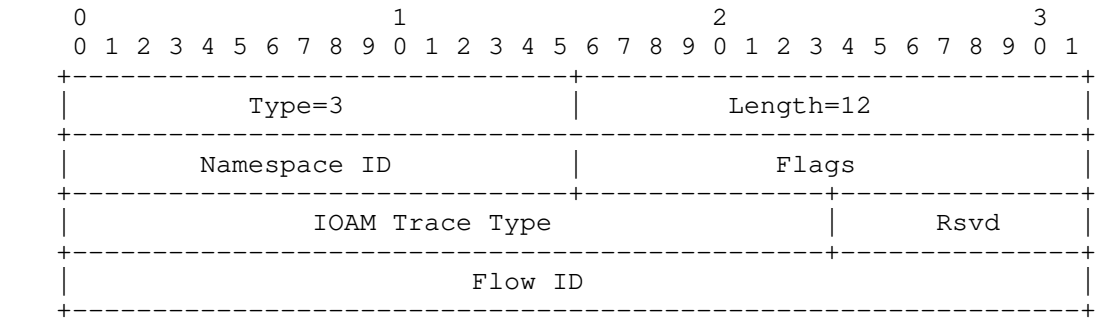


Fig. 6 IOAM Directly Export Option Sub-TLV

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:

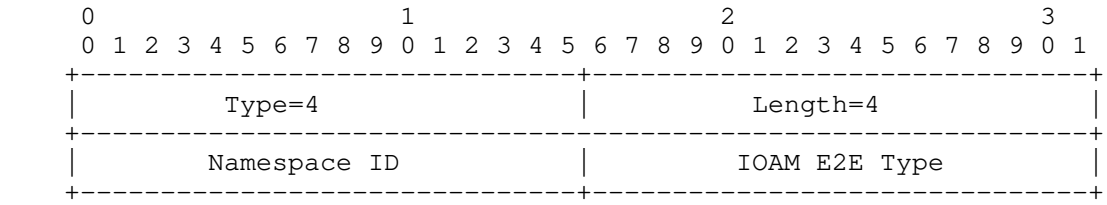


Fig. 7 IOAM Edge-to-Edge Option Sub-TLV

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:

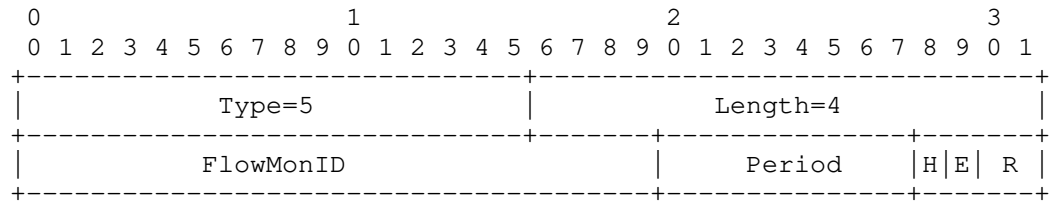


Fig. 8 Enhanced Alternate Marking Sub-TLV

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.

9. Contributors

The following people provided relevant contributions to this document:

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-02 (work in progress), October 2020.

[I-D.ietf-ippm-ioam-data]
Brockners, F., Bhandari, S., and T. Mizrahi, "Data Fields for In-situ OAM", draft-ietf-ippm-ioam-data-11 (work in progress), November 2020.

[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-02 (work in progress), November 2020.

[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-03 (work in progress), October 2020.

[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-04 (work in progress), November 2020.

[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>.

11.2. Informative References

[I-D.ietf-pce-segment-routing-ipv6] Li, C., Negi, M., Sivabalan, S., Koldychev, M., Kaladharan, P., and Y. Zhu, "PCEP Extensions for Segment Routing leveraging the IPv6 data plane", draft-ietf-pce-segment-routing-ipv6-08 (work in progress), November 2020.

[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-02 (work in progress), January 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-09 (work in progress), November 2020.

[I-D.koldychev-pce-multipath]
Koldychev, M., Sivabalan, S., Saad, T., Beeram, V., Bidgoli, H., Yadav, B., and S. Peng, "PCEP Extensions for Signaling Multipath Information", draft-koldychev-pce-multipath-04 (work in progress), October 2020.

[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

Huanan Chen
China Telecom
Guangzhou
China

Email: chenhuan6@chinatelecom.cn

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

Email: yuanhang@unionpay.com

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

Email: zhoutianran@huawei.com

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

Email: poly.li@huawei.com

Giuseppe Fioccola
Huawei
Riesstrasse, 25
Munich
Germany

Email: giuseppe.fioccola@huawei.com

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

Email: wangyali11@huawei.com

Networking Working Group
Internet-Draft
Intended status: Standards Track
Expires: August 26, 2021

Ran. Chen
ZTE Corporation
Samuel. Sidor
Cisco Systems, Inc.
Zhu. Chun
ZTE Corporation
Alex. Tokar
Mike. Koldychev
Cisco Systems, Inc.

February 22, 2021

PCEP Extensions for sid verification for SR-MPLS
draft-chen-pce-sr-mpls-sid-verification-01

Abstract

This document updates [RFC8664] to clarify usage of "SID verification" bit signalled in Path Computation Element Protocol (PCEP), and this document proposes to define a new flag for indicating the headend is explicitly requested to verify SID(s) by the PCE.

Status of This Memo

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

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

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

This Internet-Draft will expire on August 26, 2021.

Table of Contents

1. Introduction

[I-D.ietf-spring-segment-routing-policy] describes the "SID verification" bit usage. SID verification is performed when the headend is explicitly requested to verify SID(s) by the controller via the signaling protocol used. Implementations MAY provide a local configuration option to enable verification on a global or per policy or per candidate path basis.

[RFC8664] specifies 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.

This document updates [RFC8664] to clarify usage of "SID verification" bit signalled in Path Computation Element Protocol (PCEP), and this document proposes to define a new flag for indicating the headend is explicitly requested to verify SID(s) by the PCE.

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. SID verification flag(V-Flag)

3.1. Extended V-Flag in SR-ERO Subobject

Section 4.3.1 in Path Computation Element Communication Protocol (PCEP) Extensions for Segment Routing [RFC8664] describes a new ERO subobject referred to as the "SR-ERO subobject" to carry a SID and/or NAI information. A new flag is proposed in this doucument in the SR-ERO Subobject for indicating the pcc is explicitly requested to verify SID(s) by the PCE.

The format of the SR-ERO subobject as defined in [RFC8664] is:

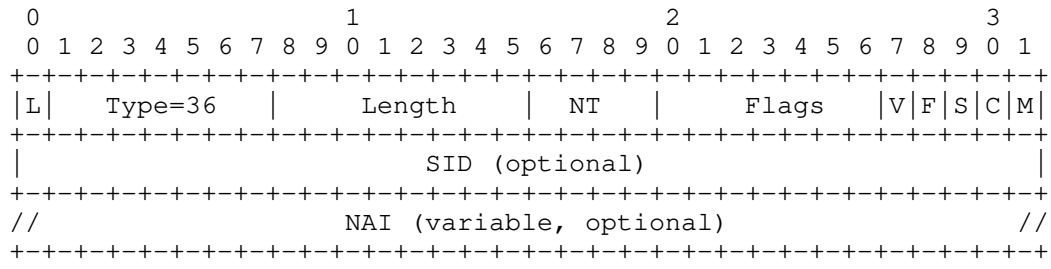


Figure 1 SR-ERO Subobject Format of extended V-Flag

V: When the V-Flag is set then PCC MUST consider the "SID verification" as described in Section 5.1 in [I-D.ietf-spring-segment-routing-policy] .

3.2. Extended V-Flag in SR-RRO Subobject

The format of the SR-RRO subobject is the same as that of the SR-ERO subobject, but without the L-Flag, per [RFC8664].

The V flag has no meaning in the SR-RRO and is ignored on receipt at the PCE.

4. Security Considerations

TBD.

5. IANA Considerations

5.1. SR-ERO Subobject

This document defines a new bit value in the sub-registry "SR-ERO Flag Field" in the "Path Computation Element Protocol (PCEP) Numbers" registry.

Bit	Name	Reference
7	SID verification(V)	This document

6. Normative references

[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-09 (work in progress), November 2020.

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

Authors' Addresses

Ran Chen
ZTE Corporation

Email: chen.ran@zte.com.cn

Samuel Sidor
Cisco Systems, Inc.

Email: ssidor@cisco.com

Chun Zhu
ZTE Corporation

Email: zhu.chun1@zte.com.cn

Alex Tokar
Cisco Systems, Inc.

Email: atokar@cisco.com

Mike Koldychev
Cisco Systems, Inc.

Email: mkoldych@cisco.com

PCE Working Group
Internet-Draft
Intended status: Experimental
Expires: August 24, 2021

D. Dhody
S. Peng
Huawei Technologies
Y. Lee
Samsung Electronics
D. Ceccarelli
Ericsson
A. Wang
China Telecom
G. Mishra
Verizon Inc.

February 20, 2021

PCEP extensions for Distribution of Link-State and TE Information
draft-dhodylee-pce-pcep-ls-20

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/.

Copyright Notice

Table of Contents

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 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.

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.

When the results of implementation and deployment are available, this document 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.

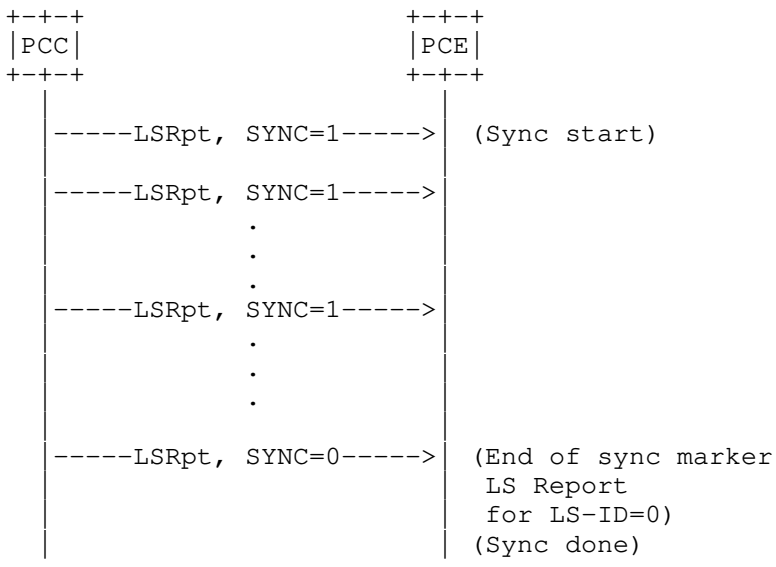


Figure 1: Successful LS synchronization

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

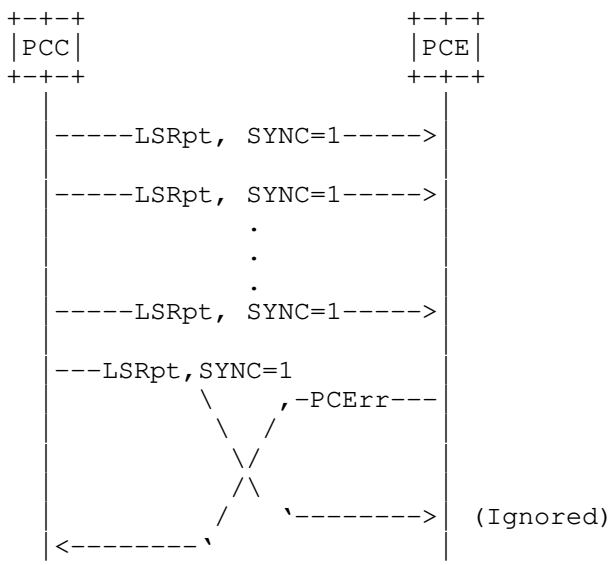


Figure 2: Failed LS synchronization (PCE failure)

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

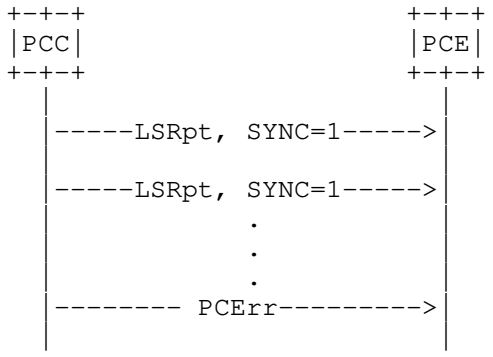


Figure 3: Failed LS synchronization (PCC failure)

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

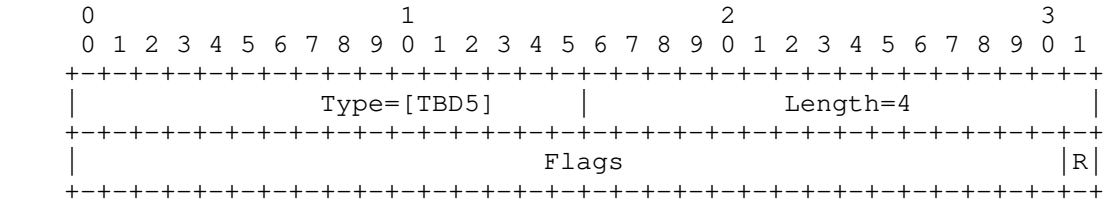
The TLV and the sub-TLV format (and padding) in this document, is as per section 7.1 of [RFC5440].

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

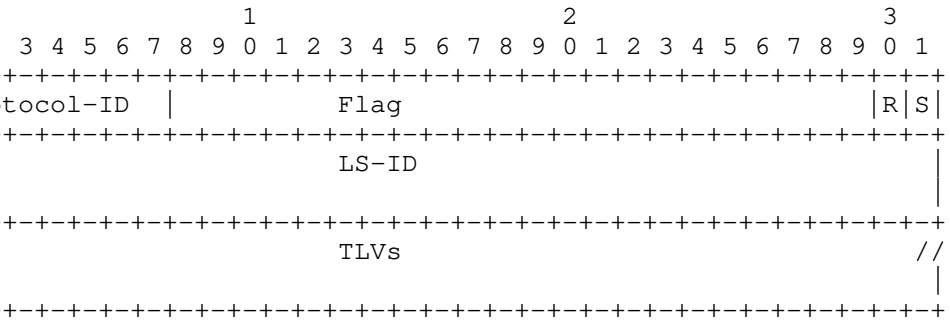
The LS (link-state) object MUST be carried within LSRpt messages and MAY be carried within PCErr messages. The LS object contains a set of fields used to specify the target node or link. It also contains a flag indicating to a PCE that the LS synchronization is in progress. The TLVs used with the LS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

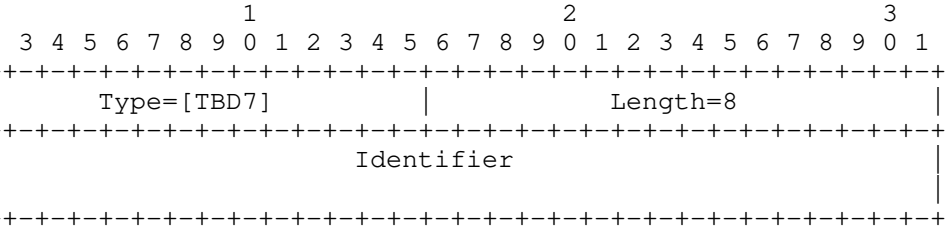
TLVs that may be included in the LS Object are described in the following sections.

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 seperate 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.

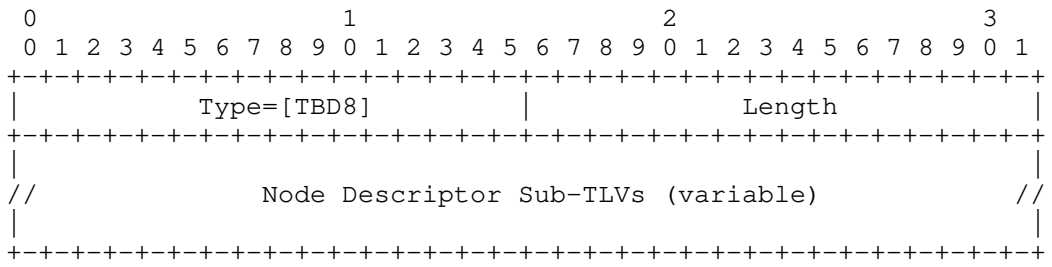
9.3.4. Local Node Descriptors TLV

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.

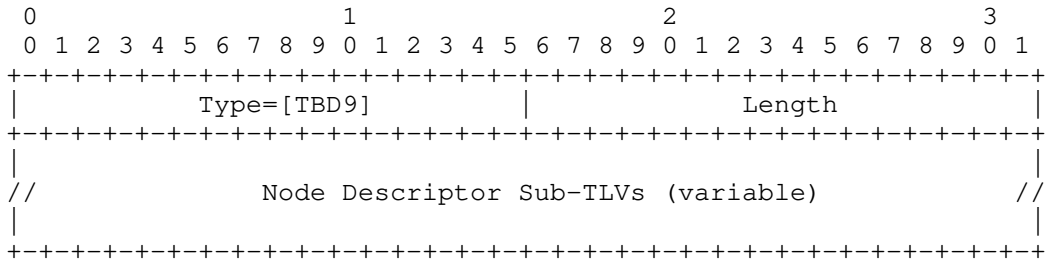


The value contains one or more Node Descriptor Sub-TLVs defined in Section 9.3.6.

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]

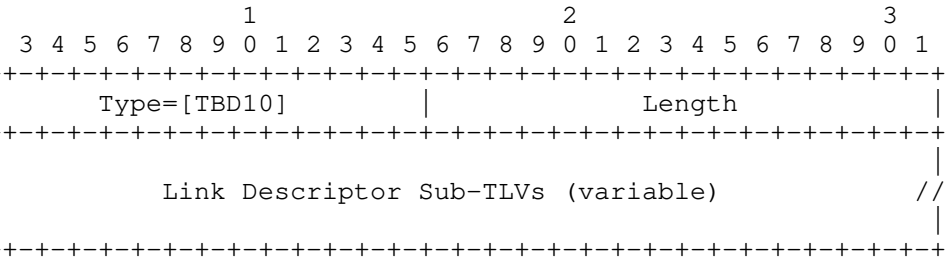
A new sub-TLV type (1) is allocated for SPEAKER-ENTITY-ID sub-TLV. The length and value fields are as per [RFC8232].

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.

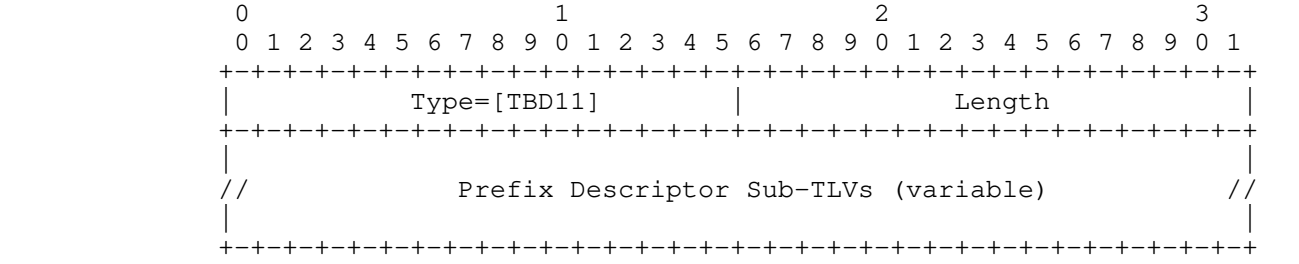
The format and semantics of the 'value' fields in most 'Link Descriptor' sub-TLVs correspond to the format and semantics of value fields in IS-IS Extended IS Reachability sub-TLVs, defined in [RFC5305], [RFC5307] and [RFC6119]. Although the encodings for 'Link Descriptor' TLVs were originally defined 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.

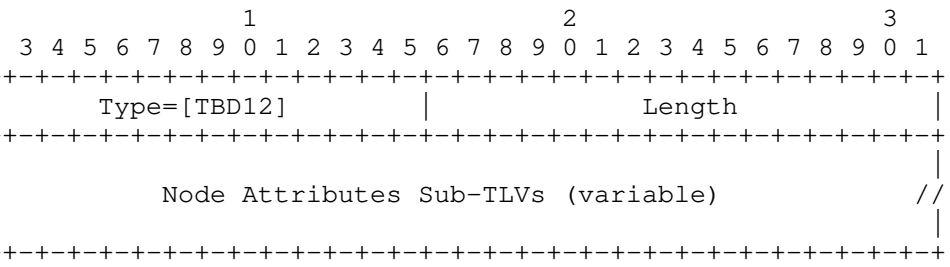


All Prefix Descriptors TLVs defined for BGP-LS can then be used with PCEP-LS as well. No new PCEP sub-TLVs for Prefix Descriptor are defined in this document.

9.3.9. PCEP-LS Attributes

9.3.9.1. Node Attributes TLV

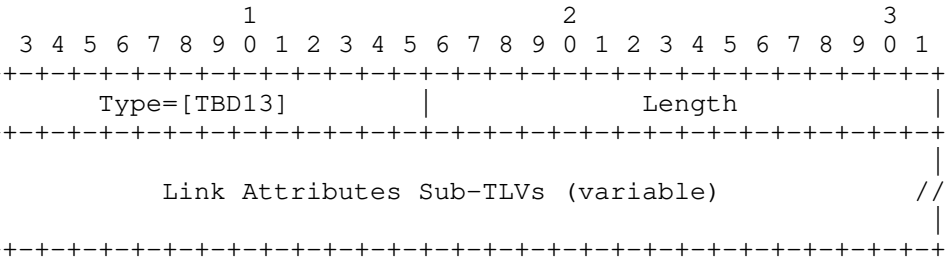
This is an optional attribute that is used to carry node attributes. This TLV is applicable for LS Node Object-Type.



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

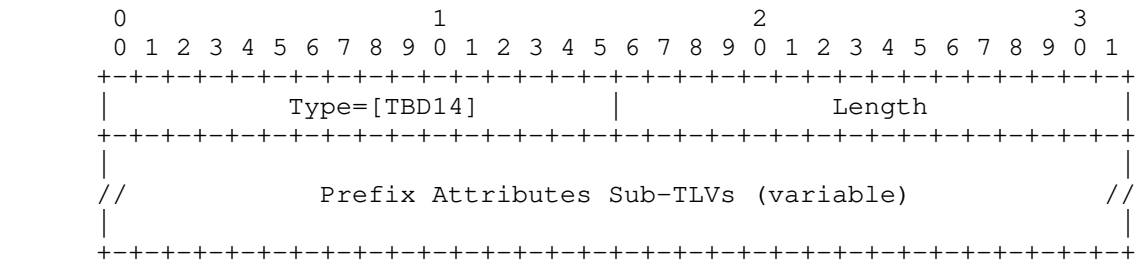
This TLV is applicable for LS Link Object-Type. The format and semantics of the 'value' fields in some 'Link Attribute' sub-TLVs correspond to the format and semantics of the 'value' fields in IS-IS Extended IS Reachability sub-TLVs, defined in [RFC5305], [RFC5307] and [I-D.ietf-idr-rfc7752bis]. Although 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

o 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 Object-Type=1 (LS Node) Object-Type=2 (LS Link) Object-Type=3 (LS IPv4 Prefix) Object-Type=4 (LS IPv6 Prefix)	[This I-D]

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

o 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

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> and the code at <ONOS-PCEP-GITHUB>.

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", draft-ietf-idr-rfc7752bis-05 (work in progress), November 2020.

[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", draft-ietf-pce-pcep-flowspec-12 (work in progress), October 2020.

[I-D.ietf-pce-pcep-yang]
Dhody, D., Hardwick, J., Beeram, V., and J. Tantsura, "A YANG Data Model for Path Computation Element Communications Protocol (PCEP)", draft-ietf-pce-pcep-yang-15 (work in progress), October 2020.

[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", draft-ietf-pce-vn-association-03 (work in progress), October 2020.

[I-D.ietf-teas-actn-requirements]
Lee, Y., Ceccarelli, D., Miyasaka, T., Shin, J., and K. Lee, "Requirements for Abstraction and Control of TE Networks", draft-ietf-teas-actn-requirements-09 (work in progress), March 2018.

[I-D.kondreddy-pce-pcep-ls-sync-optimizations]
Kondreddy, V. and M. Negi, "Optimizations of PCEP Link-State(LS) Synchronization Procedures", draft-kondreddy-pce-pcep-ls-sync-optimizations-00 (work in progress), October 2015.

[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., 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>.

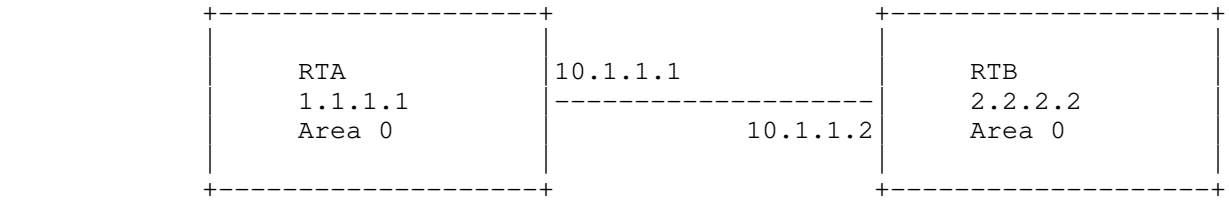
[RFC8637] Dhody, D., Lee, Y., and D. Ceccarelli, "Applicability of the Path Computation Element (PCE) to the Abstraction and Control of TE Networks (ACTN)", RFC 8637, DOI 10.17487/RFC8637, July 2019, <https://www.rfc-editor.org/info/rfc8637>.

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: Router-ID: 1.1.1.1
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: Router-ID: 1.1.1.1
TLV - Remote Node Descriptors
Sub-TLV - 514: OSPF Area-ID: 0.0.0.0
Sub-TLV - 515: Router-ID: 2.2.2.2
TLV - Link Descriptors
Sub-TLV - 259: IPv4 interface: 10.1.1.1
Sub-TLV - 260: IPv4 neighbor: 10.1.1.2
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: Router-ID: 2.2.2.2

```
      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: Router-ID: 2.2.2.2
      TLV - Remote Node Descriptors
          Sub-TLV - 514: OSPF Area-ID: 0.0.0.0
          Sub-TLV - 515: Router-ID: 1.1.1.1
      TLV - Link Descriptors
          Sub-TLV - 259: IPv4 interface: 10.1.1.2
          Sub-TLV - 260: IPv4 neighbor: 10.1.1.1
      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: Router-ID: 11.11.11.11 (abstract)

   LS Link
      TLV - Local Node Descriptors
          Sub-TLV - 512: Autonomous System: 100
          Sub-TLV - 515: Router-ID: 11.11.11.11 (abstract)
      TLV - Remote Node Descriptors
          Sub-TLV - 512: Autonomous System: 200 (Domain 2)
          Sub-TLV - 515: Router-ID: 22.22.22.22 (abstract)
      TLV - Link Descriptors
          Sub-TLV - 259: IPv4 interface: 11.1.1.1
          Sub-TLV - 260: IPv4 neighbor: 11.1.1.2
      TLV - Link Attributes TLV
          Sub-TLV(s)

   LS Link
      TLV - Local Node Descriptors
          Sub-TLV - 512: Autonomous System: 100
          Sub-TLV - 515: Router-ID: 11.11.11.11 (abstract)
      TLV - Remote Node Descriptors
          Sub-TLV - 512: Autonomous System: 200
          Sub-TLV - 515: Router-ID: 22.22.22.22 (abstract)
      TLV - Link Descriptors
          Sub-TLV - 259: IPv4 interface: 12.1.1.1
          Sub-TLV - 260: IPv4 neighbor: 12.1.1.2
      TLV - Link Attributes TLV
          Sub-TLV(s)

   LS Link
      TLV - Local Node Descriptors
          Sub-TLV - 512: Autonomous System: 100
          Sub-TLV - 515: Router-ID: 11.11.11.11 (abstract)
      TLV - Remote Node Descriptors
          Sub-TLV - 512: Autonomous System: 400 (Domain 4)
          Sub-TLV - 515: Router-ID: 44.44.44.44 (abstract)
      TLV - Link Descriptors
          Sub-TLV - 259: IPv4 interface: 13.1.1.1
          Sub-TLV - 260: IPv4 neighbor: 13.1.1.2
      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, Karnataka 560066
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, Beijing 102209
China

EMail: wangaj3@chinatelecom.cn

Gyan Mishra
Verizon Inc.

EMail: gyan.s.mishra@verizon.com

PCE Working Group
Internet-Draft
Intended status: Standards Track
Expires: August 11, 2021

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

February 7, 2021

PCEP Extension for Native IP Network
draft-ietf-pce-pcep-extension-native-ip-11

Abstract

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

Table of Contents

1. Introduction

Generally, Multiprotocol Label Switching Traffic Engineering (MPLS-TE) requires the corresponding network devices support Multiprotocol Label Switching (MPLS) or Resource ReSerVation Protocol (RSVP)/Label Distribution Protocol (LDP) technologies to assure the End-to-End (E2E) traffic performance. But in native IP network, there will be no such signaling protocol to synchronize the action among different network devices. It is necessary to use the central control mode that described in [RFC8283] to correlate the forwarding behavior among different network devices. Draft [I-D.ietf-teas-pce-native-ip] describes the architecture and solution philosophy for the E2E traffic assurance in Native IP network via Multi Border Gateway Protocol (BGP) solution. This draft describes the corresponding Path Computation Element Communication Protocol (PCEP) extensions to transfer the key information about BGP peer info, peer prefix association and the explicit peer route on on-path routers.

2. Conventions used in this document

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

3. Terminology

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

The following terms are defined in this document:

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

4. Capability Advertisemnt

4.1. Open message

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

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

- PST = TBD1: Path is a Native IP path as per [I-D.ietf-teas-pce-native-ip].

A PCEP speaker MUST indicate its support of the function described in this document by sending a PATH-SETUP-TYPE-CAPABILITY TLV in the OPEN object with this new PST included in the PST list.

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

N (NATIVE-IP-TE-CAPABILITY - 1 bit - TBD2): If set to 1 by a PCEP speaker, it indicates that the PCEP speaker is capable for TE in Native IP network as specified in this document. The flag MUST be set by both the PCC and PCE in order to support this extension.

If a PCEP speaker receives the PATH-SETUP-TYPE-CAPABILITY TLV with the newly defined path setup type, but without the N bit set in PCECC-CAPABILITY sub-TLV, it MUST:

- Send a PCErr message with Error-Type=10(Reception of an invalid object) and Error-Value TBD3(PCECC NATIVE-IP-TE-CAPABILITY bit is not set).

- Terminate the PCEP session

5. PCEP messages

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

The extended PCInitiate message described in [I-D.ietf-pce-pcep-extension-for-pce-controller] is used to download

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

5.1. The PCInitiate message

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

The format of the extended PCInitiate message is as follows:

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

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

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

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

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

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

    The LSP and SRP object is defined in [RFC8231].
```

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

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

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

5.2. The PCRpt message

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

The format of the PCRpt message is as follows:

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

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

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

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

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

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

The error handling for missing CCI object is as per [I-D.ietf-pce-pcep-extension-for-pce-controller]. Further only one of BPI, EPR, or PPA object MUST be present.

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

6. PCECC Native IP TE Procedures

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

6.1. BGP Session Establishment Procedures

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

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

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

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

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

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

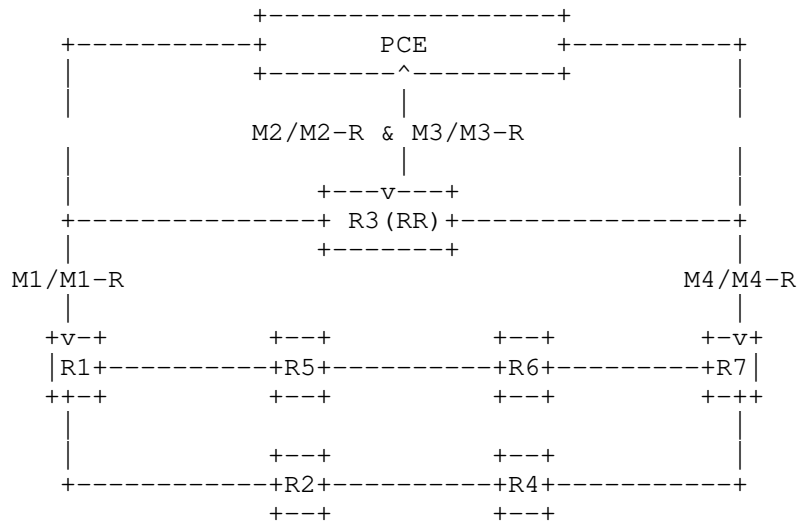


Figure 1: BGP Session Establishment Procedures(R3 act as RR)

The message number, message peers, message type and message key parameters in the above figures are shown in below table:

Table 1: Message Information

No.	Peers	Type	Message Key Parameters
M1 M1-R	PCE/R1	PCInitiate PCRpt	CC-ID=X1(Symbolic Path Name=Class A) BPI Object(Local_IP=R1_A,Peer_IP=R3_A)
M2 M2-R	PCE/R3	PCInitiate PCRpt	CC-ID=X2(Symbolic Path Name=Class A) BPI Object(Local_IP=R3_A,Peer_IP=R1_A)
M3 M3-R	PCE/R3	PCInitiate PCRpt	CC-ID=X3(Symbolic Path Name=Class A) BPI Object(Local_IP=R3_A,Peer_IP=R7_A)
M4 M4-R	PCE/R7	PCInitiate PCRpt	CC-ID=X4(Symbolic Path Name=Class A) BPI Object(Local_IP=R7_A,Peer_IP=R3_A)

If the PCC cannot establish the BGP session that required by this object, it should report the error values via PCErr message with the newly defined error type(Error-type=TBD6) and error value(Error-value=TBD7, Peer AS not match; or Error-Value=TBD8, Peer IP can't be reached), which is indicated in Section 9

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

6.2. Explicit Route Establish Procedures

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

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

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

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

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

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

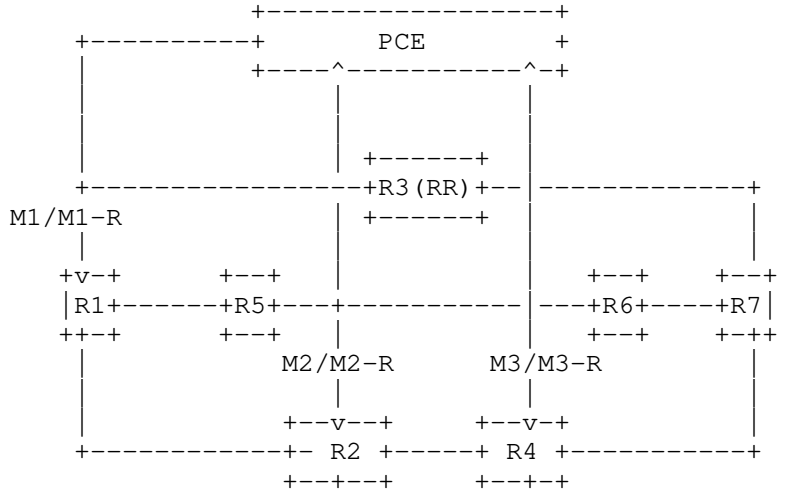


Figure 2: Explicit Route Establish Procedures(From R1 to R7)

The message number, message peers, message type and message key parameters in the above figures are shown in below table:

Table 2: Message Information

No.	Peers	Type	Message Key Parameters
M1 M1-R	PCE/R1	PCInitiate PCRpt	CC-ID=X1(Symbolic Path Name=Class A) EPR Object(Peer Address=R7_A,Next Hop=R2_A)
M2 M2-R	PCE/R2	PCInitiate PCRpt	CC-ID=X2(Symbolic Path Name=Class A) EPR Object(Peer Address=R7_A,Next Hop=R4_A)
M3 M3-R	PCE/R4	PCInitiate PCRpt	CC-ID=X3(Symbolic Path Name=Class A) EPR Object(Peer Address=R7_A,Next Hop=R7_A)

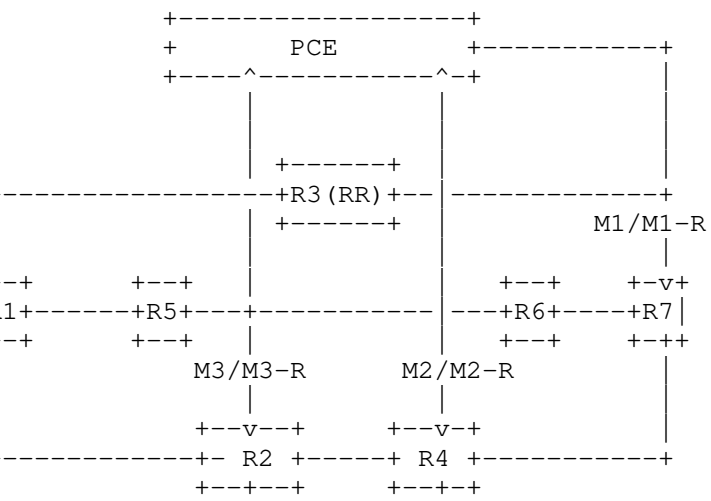


Figure 3: Explicit Route Establish Procedures(From R7 to R1)

The message number, message peers, message type and message key parameters in the above figures are shown in below table:

Table 3: Message Information

No.	Peers	Type	Message Key Parameters
M1 M1-R	PCE/R7	PCInitiate PCRpt	CC-ID=X1(Symbolic Path Name=Class A) EPR Object(Peer Address=R1_A,Next Hop=R4_A)
M2 M2-R	PCE/R4	PCInitiate PCRpt	CC-ID=X2(Symbolic Path Name=Class A) EPR Object(Peer Address=R1_A,Next Hop=R2_A)
M3 M3-R	PCE/R2	PCInitiate PCRpt	CC-ID=X3(Symbolic Path Name=Class A) EPR Object(Peer Address=R1_A,Next Hop=R1_A)

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

Upon the error occurs, the PCC SHOULD send the corresponding error via PCErr message, with an error information (Error-type=TBD6, Error-value=TBD12, Explicit Peer Route Error) that defined in Section 9.

When the peer info that associated with the Symbolic Path Name is not the same as the peer info that indicated in BPI object in PCC, an

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

6.3. BGP Prefix Advertisement Procedures

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

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

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

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

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

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

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

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

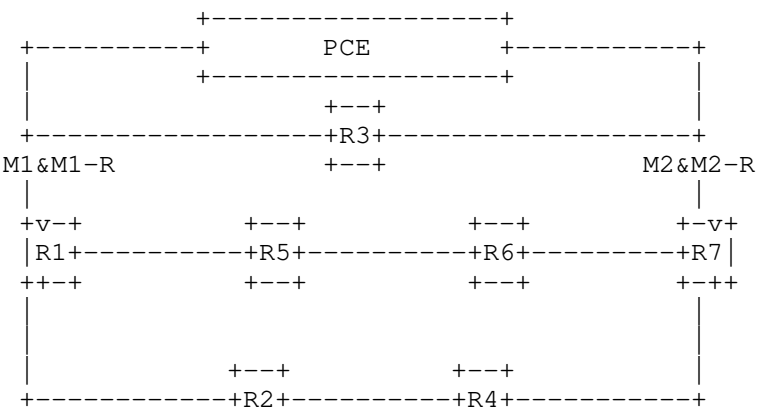


Figure 4: BGP Prefix Advertisement Procedures

Table 4: Message Information

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

7. New PCEP Objects

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

7.1. CCI Object

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

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

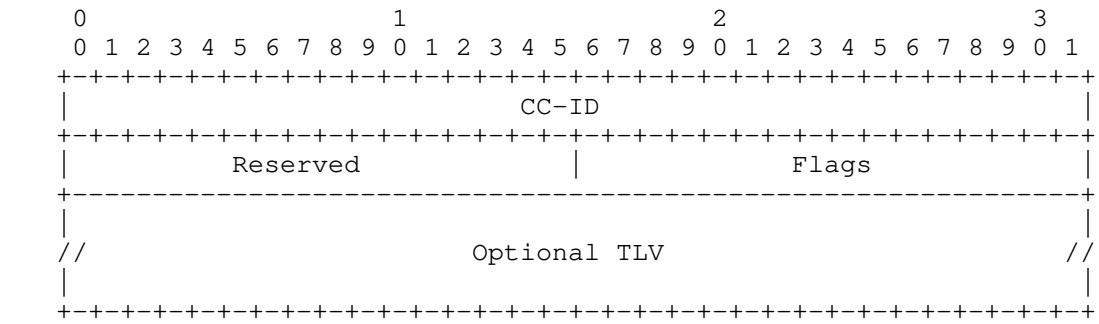


Figure 5: CCI Object for Native IP

Figure 1

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

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

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

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

7.2. BGP Peer Info Object

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

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

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

BGP Peer Info Object-Class is TBD14

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

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

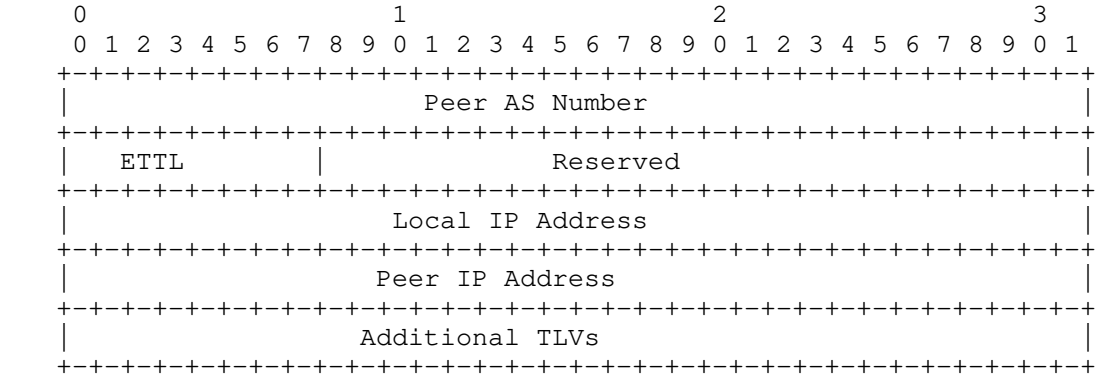


Figure 6: BGP Peer Info Object Body Format for IPv4

The format of the BGP Peer Info object body for IPv6(Object-Type=2) is as follows:

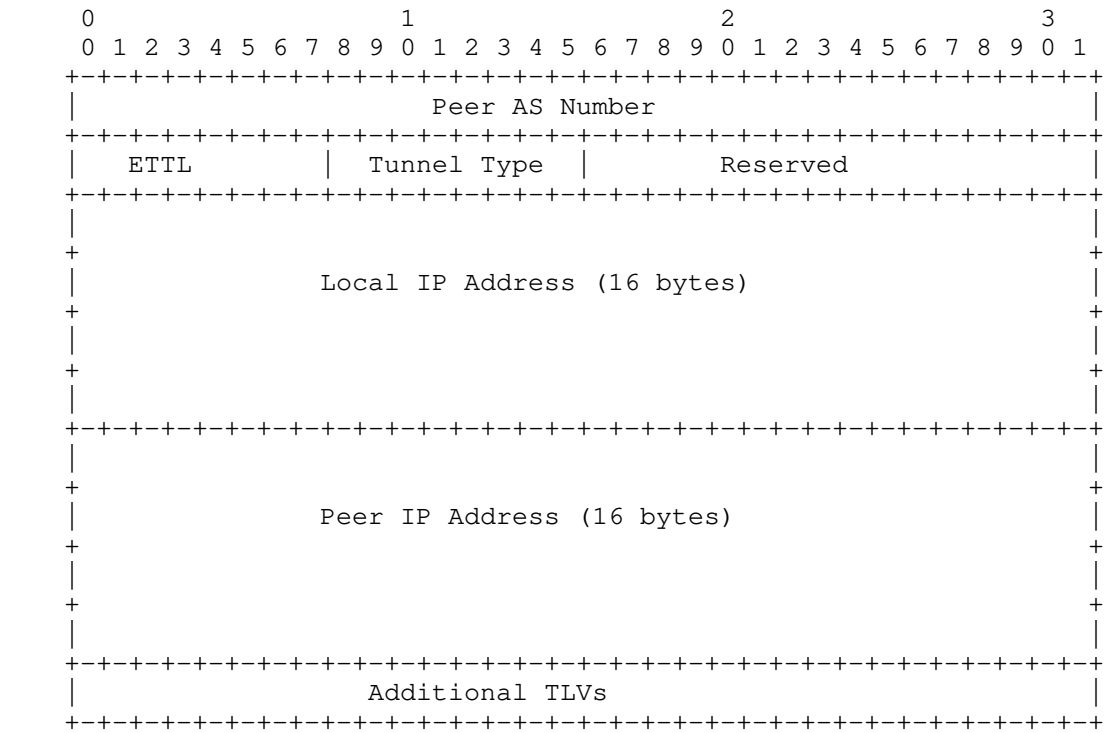


Figure 7: BGP Peer Info Object Body Format for IPv6

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

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

Tunnel Type: 1 Byte, indicate the tunnel type that used to transfer the traffic that identified by the prefixes that advertised by the corresponding BGP peer. Value 0 indicate no tunnel is used. Other value can refer to the IANA allocation value in "BGP Tunnel Encapsulation Attribute Tunnel Types".

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

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

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

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

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

7.3. Explicit Peer Route Object

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

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

Explicit Peer Route Object-Class is TBD15.

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

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

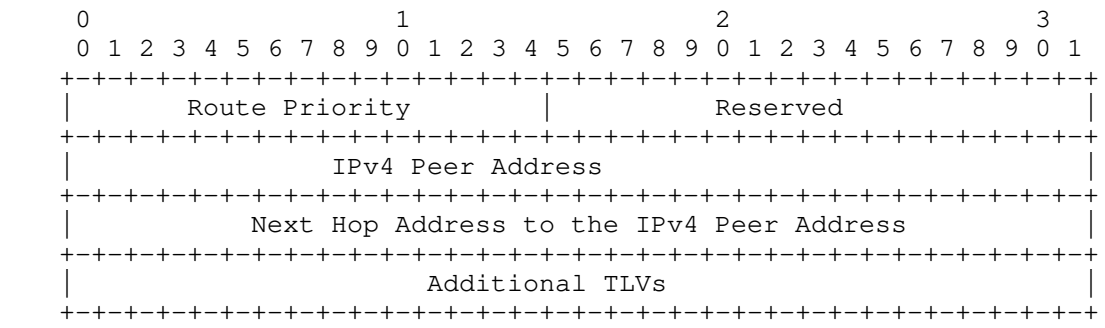


Figure 8: Explicit Peer Route Object Body Format for IPv4

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

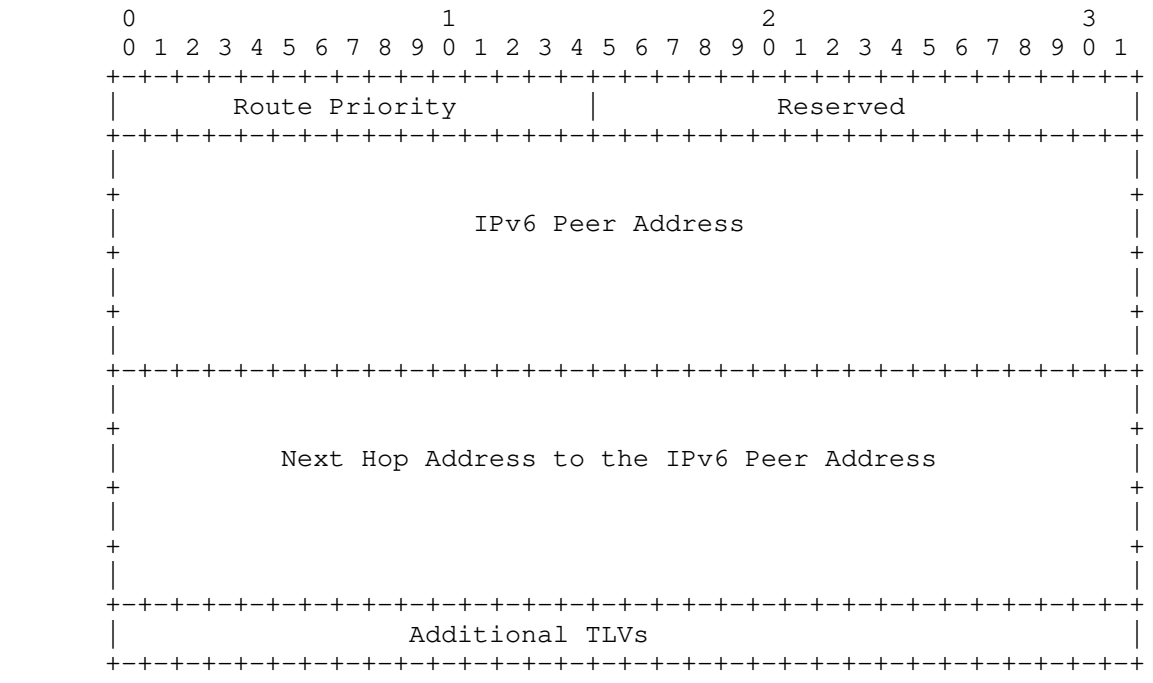


Figure 9: Explicit Peer Route Object Body Format for IPv6

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

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

Peer Address: To indicate the peer address.

Next Hop Address to the Peer: To indicate the next hop address to the corresponding peer.

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

7.4. Peer Prefix Association Object

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

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

Peer Prefix Association Object-Class is TBD16

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

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

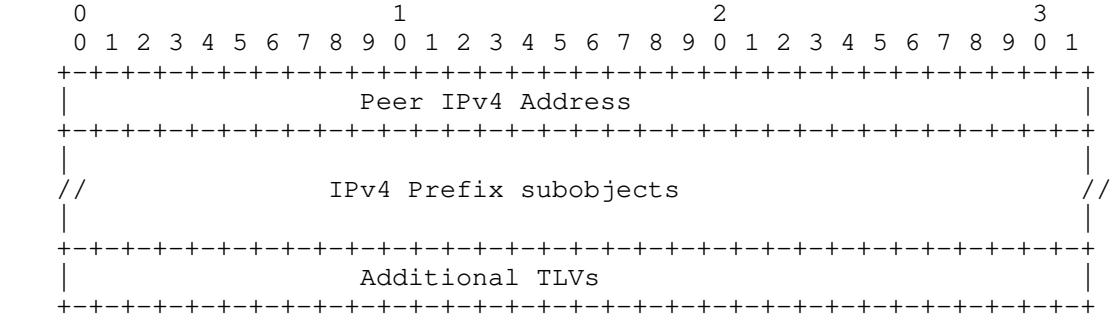


Figure 10: Peer Prefix Association Object Body Format for IPv4

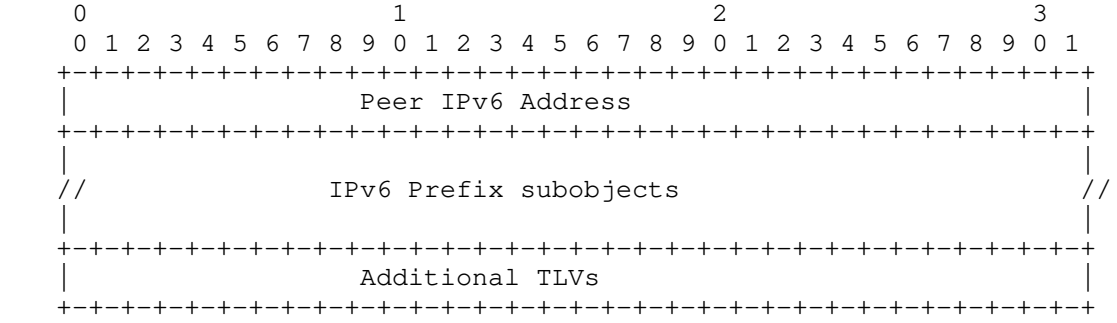


Figure 11: Peer Prefix Association Object Body Format for IPv6

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

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

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

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

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

8. End to End Path Protection

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

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

9. New Error-Types and Error-Values Defined

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

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

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

10. Deployment Considerations

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

[RFC8232] describes the state synchronization procedure between stateful PCE and PCC. The communication of PCE and PCC described in this draft should also follow this procedures, treat the three newly

defined objects that associated with the same symbolic path name as the attribute of the same path in the LSP-DB.

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

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

11. Security Considerations

Service provider should consider the protection of PCE and their communication with the underlay devices, which is described in document [RFC5440] and [RFC8253]

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	Native IP TE Path	This document

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

[I-D.ietf-pce-pcep-extension-for-pce-controller] created a sub-registry within the "Path Computation Element Protocol (PCEP) Numbers" registry to manage the value of the PCECC-CAPABILITY sub-TLV's 32-bits Flag field. IANA is requested to allocate a new bit position within this registry, as follows:

Value	Description	Reference
TBD2(N)	NATIVE-IP-TE-CAPABILITY	This document

12.3. PCEP Object Types

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

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

12.4. PCEP-Error Object

IANA is requested to allocate new error types and error values within the "PCEP-ERROR Object Error Types and Values" sub-registry of the PCEP Numbers registry for the following errors::

Error-Type	Meaning	Error-value	Reference
6	Mandatory Object missing	TBD4:Native IP object missing	This document
10	Reception of an invalid object	TBD3:PCECC NATIVE-IP-TE-CAPABILITY bit is not set	This document
19	Invalid Operation	TBD5:Only one of the BPI,EPR or PPA object can be included in this message	This document
TBD6	Native IP TE failure	TBD7:Peer AS not match TBD8:Peer IP can't be reached TBD9:Local IP is in use TBD10:Remote IP is in use TBD11:Exist BGP session broken TBD12:Explicit Peer Route Error TBD17:EPR/BPI Peer Info mismatch TBD18:BPI/PPA Address Family mismatch TBD19:PPA/BPI Peer Info mismatch	This document

13. Contributor

Dhruv Dhody has contributed the contents of this draft.

14. Acknowledgement

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

15. Normative References

[I-D.ietf-pce-pcep-extension-for-pce-controller]
Li, Z., Peng, S., Negi, M., Zhao, Q., and C. Zhou, "PCEP Procedures and Protocol Extensions for Using PCE as a Central Controller (PCECC) of LSPs", draft-ietf-pce-pcep-extension-for-pce-controller-10 (work in progress), January 2021.

[I-D.ietf-teas-pce-native-ip]
Wang, A., Khasanov, B., Zhao, Q., and H. Chen, "Path Computation Element (PCE) based Traffic Engineering (TE) in Native IP Networks", draft-ietf-teas-pce-native-ip-16 (work in progress), January 2021.

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

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

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

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

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

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

[RFC8253] Lopez, D., Gonzalez de Dios, O., Wu, Q., and D. Dhody, "PCEPS: Usage of TLS to Provide a Secure Transport for the Path Computation Element Communication Protocol (PCEP)", RFC 8253, DOI 10.17487/RFC8253, October 2017, <https://www.rfc-editor.org/info/rfc8253>.

[RFC8281] Crabbe, E., Minei, I., Sivabalan, S., and R. Varga, "Path Computation Element Communication Protocol (PCEP) Extensions for PCE-Initiated LSP Setup in a Stateful PCE Model", RFC 8281, DOI 10.17487/RFC8281, December 2017, <https://www.rfc-editor.org/info/rfc8281>.

[RFC8283] Farrel, A., Ed., Zhao, Q., Ed., Li, Z., and C. Zhou, "An Architecture for Use of PCE and the PCE Communication Protocol (PCEP) in a Network with Central Control", RFC 8283, DOI 10.17487/RFC8283, December 2017, <https://www.rfc-editor.org/info/rfc8283>.

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

[RFC8697] Minei, I., Crabbe, E., Sivabalan, S., Ananthakrishnan, H., Dhody, D., and Y. Tanaka, "Path Computation Element Communication Protocol (PCEP) Extensions for Establishing Relationships between Sets of Label Switched Paths (LSPs)", RFC 8697, DOI 10.17487/RFC8697, January 2020, <https://www.rfc-editor.org/info/rfc8697>.

[RFC8735] Wang, A., Huang, X., Kou, C., Li, Z., and P. Mi, "Scenarios and Simulation Results of PCE in a Native IP Network", RFC 8735, DOI 10.17487/RFC8735, February 2020, <https://www.rfc-editor.org/info/rfc8735>.

Authors' Addresses

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

Email: wangaj3@chinatelecom.cn

Boris Khasanov
Yandex LLC
Ulitsa Lva Tolstogo 16
Moscow
Russia

Email: bhassanov@yahoo.com

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

Email: fsheng@huawei.com

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

Email: tanren@huawei.com

Chun Zhu
ZTE Corporation
50 Software Avenue, Yuhua District
Nanjing, Jiangsu 210012
China

Email: zhu.chun1@zte.com.cn

PCE Working Group
Internet-Draft
Intended status: Standards Track
Expires: September 27, 2021

A. Wang
China Telecom
B. Khasanov
Yandex LLC
S. Fang
R. Tan
Huawei Technologies,Co.,Ltd
C. Zhu
ZTE Corporation
March 26, 2021

PCEP Extension for Native IP Network
draft-ietf-pce-pcep-extension-native-ip-13

Abstract

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

Status of This Memo

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

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

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

This Internet-Draft will expire on September 27, 2021.

Table of Contents

1. Introduction

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

2. Conventions used in this document

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

3. Terminology

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

The following terms are defined in this document:

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

- o QoS: Quality of Service

4. Capability Advertisemnt

4.1. Open message

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

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

- o PST = TBD1: Path is a Native IP path as per [I-D.ietf-teas-pce-native-ip].

A PCEP speaker MUST indicate its support of the function described in this document by sending a PATH-SETUP-TYPE-CAPABILITY TLV in the OPEN object with this new PST included in the PST list.

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

N (NATIVE-IP-TE-CAPABILITY - 1 bit - TBD2): If set to 1 by a PCEP speaker, it indicates that the PCEP speaker is capable for TE in Native IP network as specified in this document. The flag MUST be set by both the PCC and PCE in order to support this extension.

If a PCEP speaker receives the PATH-SETUP-TYPE-CAPABILITY TLV with the newly defined path setup type, but without the N bit set in PCECC-CAPABILITY sub-TLV, it MUST:

- o Send a PCErr message with Error-Type=10(Reception of an invalid object) and Error-Value TBD3(PCECC NATIVE-IP-TE-CAPABILITY bit is not set).

- o Terminate the PCEP session

5. PCEP messages

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

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

5.1. The PCInitiate message

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

The format of the extended PCInitiate message is as follows:

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

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

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

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

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

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

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

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

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

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

5.2. The PCRpt message

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

The format of the PCRpt message is as follows:

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

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

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

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

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


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

The error handling for missing CCI object is as per [I-D.ietf-pce-pcep-extension-for-pce-controller]. Further only one of BPI, EPR, or PPA object MUST be present.

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

6. PCECC Native IP TE Procedures

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

6.1. BGP Session Establishment Procedures

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

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

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

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

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

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

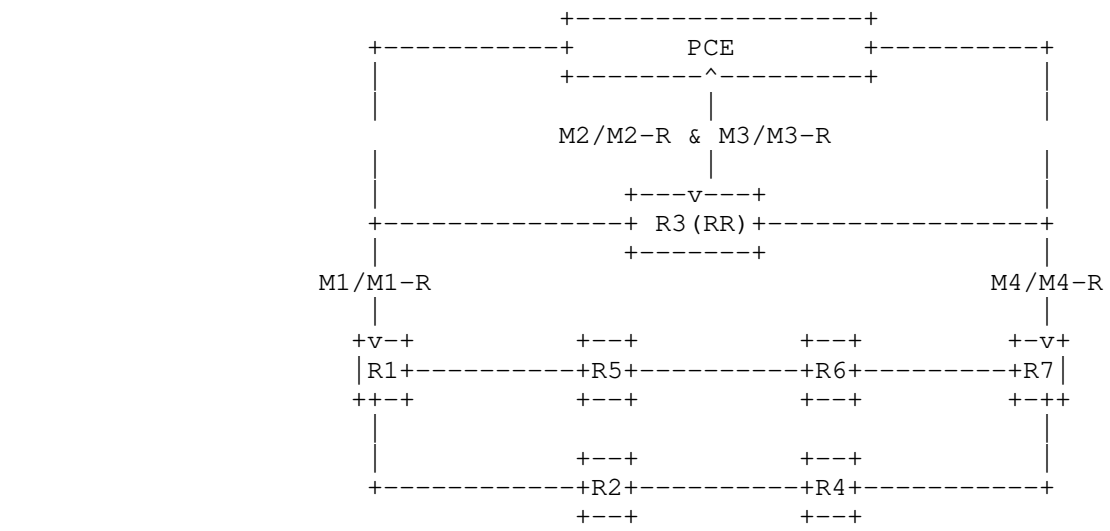


Figure 1: BGP Session Establishment Procedures(R3 act as RR)

The message number, message peers, message type and message key parameters in the above figures are shown in below table:

Table 1: Message Information

No.	Peers	Type	Message Key Parameters
M1 M1-R	PCE/R1	PCInitiate PCRpt	CC-ID=X1(Symbolic Path Name=Class A) BPI Object(Local_IP=R1_A,Peer_IP=R3_A)
M2 M2-R	PCE/R3	PCInitiate PCRpt	CC-ID=X2(Symbolic Path Name=Class A) BPI Object(Local_IP=R3_A,Peer_IP=R1_A)
M3 M3-R	PCE/R3	PCInitiate PCRpt	CC-ID=X3(Symbolic Path Name=Class A) BPI Object(Local_IP=R3_A,Peer_IP=R7_A)
M4 M4-R	PCE/R7	PCInitiate PCRpt	CC-ID=X4(Symbolic Path Name=Class A) BPI Object(Local_IP=R7_A,Peer_IP=R3_A)

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

If the Local_IP or Peer_IP within BPI object is used in other existing BGP sessions, the PCC should report such error situation via

PCErr message with Err-type=TBD6 and error value(Error-value=TBD9, Local IP is in use; Error-value=TBD10, Remote IP is in use).

6.2. Explicit Route Establish Procedures

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

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

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

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

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

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

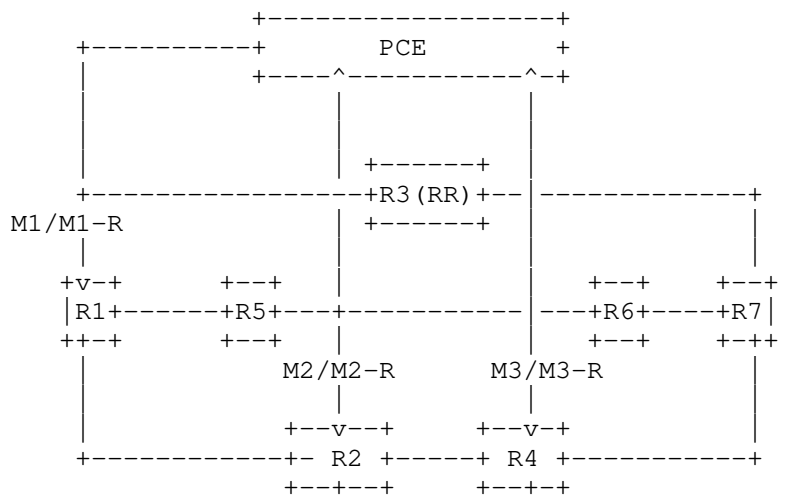


Figure 2: Explicit Route Establish Procedures(From R1 to R7)

The message number, message peers, message type and message key parameters in the above figures are shown in below table:

Table 2: Message Information

No.	Peers	Type	Message Key Parameters
M1 M1-R	PCE/R1	PCInitiate PCRpt	CC-ID=X1(Symbolic Path Name=Class A) EPR Object(Peer Address=R7_A,Next Hop=R2_A)
M2 M2-R	PCE/R2	PCInitiate PCRpt	CC-ID=X2(Symbolic Path Name=Class A) EPR Object(Peer Address=R7_A,Next Hop=R4_A)
M3 M3-R	PCE/R4	PCInitiate PCRpt	CC-ID=X3(Symbolic Path Name=Class A) EPR Object(Peer Address=R7_A,Next Hop=R7_A)

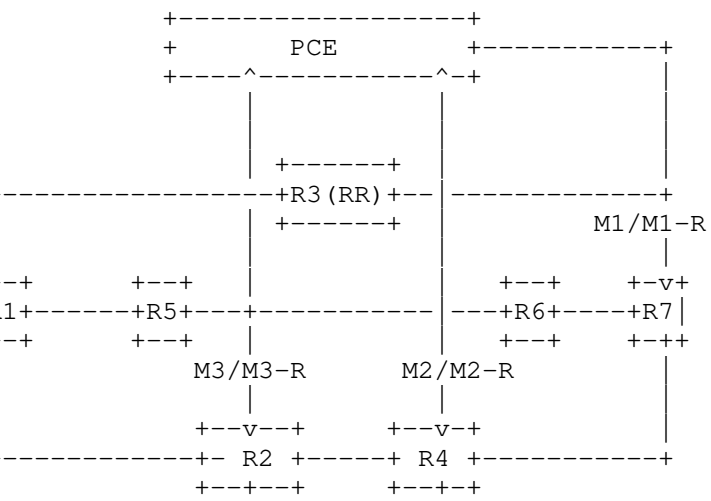


Figure 3: Explicit Route Establish Procedures(From R7 to R1)

The message number, message peers, message type and message key parameters in the above figures are shown in below table:

Table 3: Message Information

No.	Peers	Type	Message Key Parameters
M1 M1-R	PCE/R7	PCInitiate PCRpt	CC-ID=X1(Symbolic Path Name=Class A) EPR Object(Peer Address=R1_A,Next Hop=R4_A)
M2 M2-R	PCE/R4	PCInitiate PCRpt	CC-ID=X2(Symbolic Path Name=Class A) EPR Object(Peer Address=R1_A,Next Hop=R2_A)
M3 M3-R	PCE/R2	PCInitiate PCRpt	CC-ID=X3(Symbolic Path Name=Class A) EPR Object(Peer Address=R1_A,Next Hop=R1_A)

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

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

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

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

6.3. BGP Prefix Advertisement Procedures

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

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

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

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

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

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

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

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

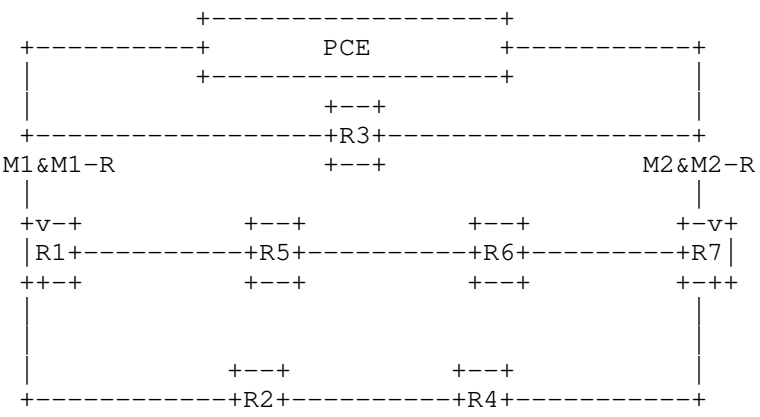


Figure 4: BGP Prefix Advertisement Procedures

Table 4: Message Information

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

7. New PCEP Objects

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

7.1. CCI Object

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

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

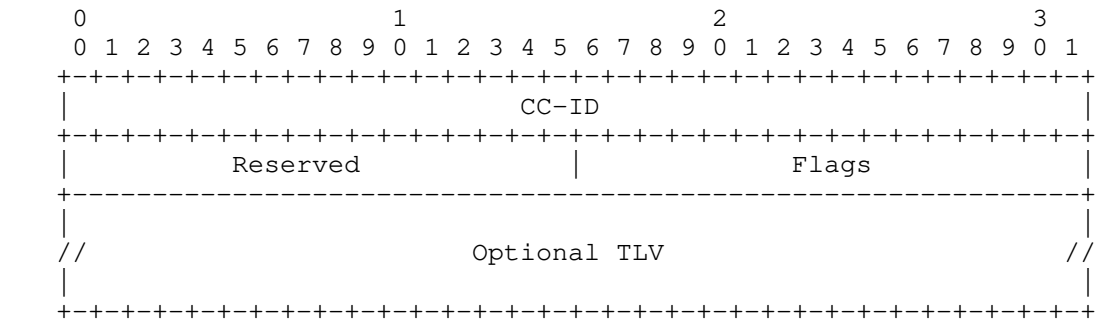


Figure 5: CCI Object for Native IP

Figure 1

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

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

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

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

7.2. BGP Peer Info Object

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

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

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

BGP Peer Info Object-Class is TBD14

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

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

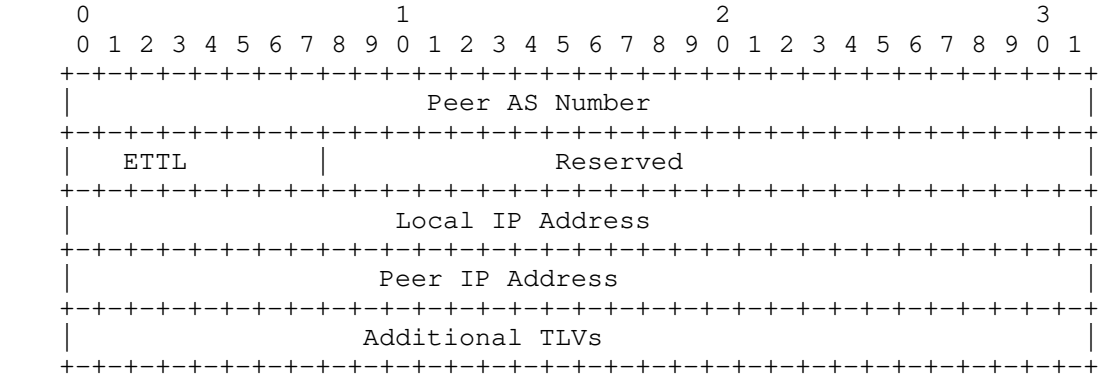


Figure 6: BGP Peer Info Object Body Format for IPv4

The format of the BGP Peer Info object body for IPv6(Object-Type=2) is as follows:

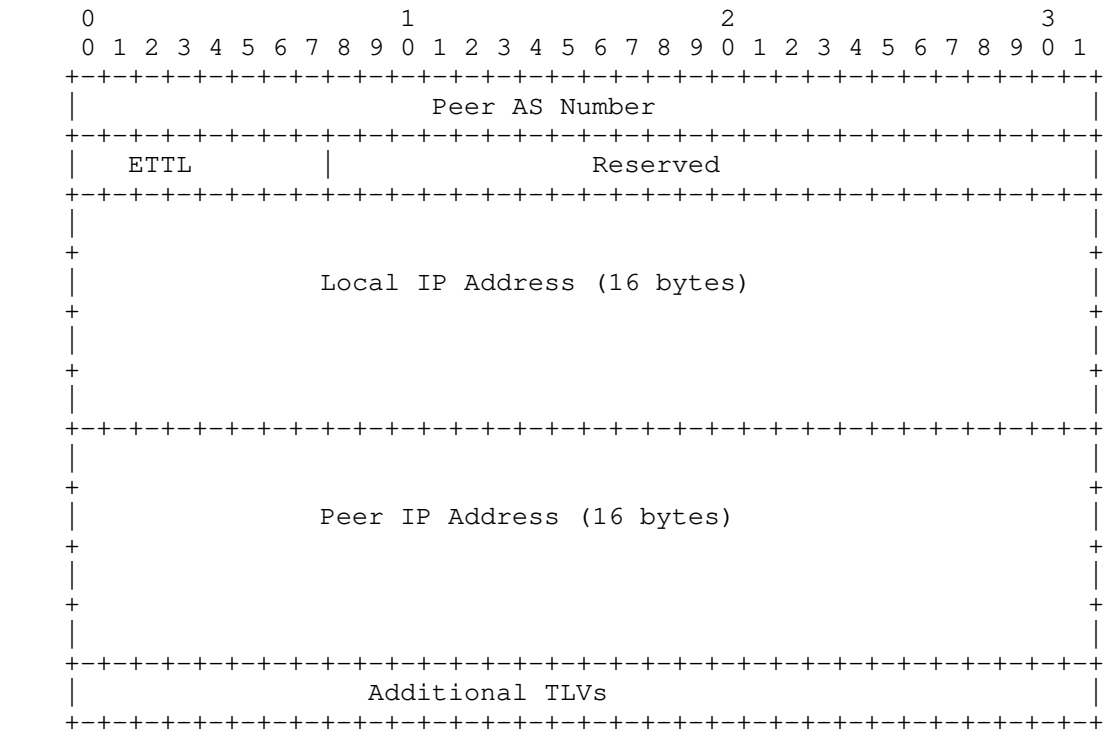


Figure 7: BGP Peer Info Object Body Format for IPv6

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

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

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

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

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

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

When PCC receives BPI object, with Object-Type=1, it should try to establish BGP session with the peer in AFI/SAFI=1/1; when PCC

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

7.3. Explicit Peer Route Object

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

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

Explicit Peer Route Object-Class is TBD15.

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

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

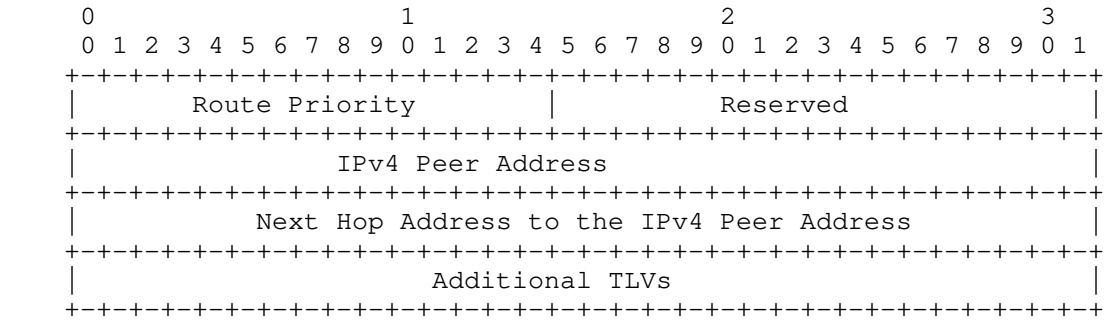


Figure 8: Explicit Peer Route Object Body Format for IPv4

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

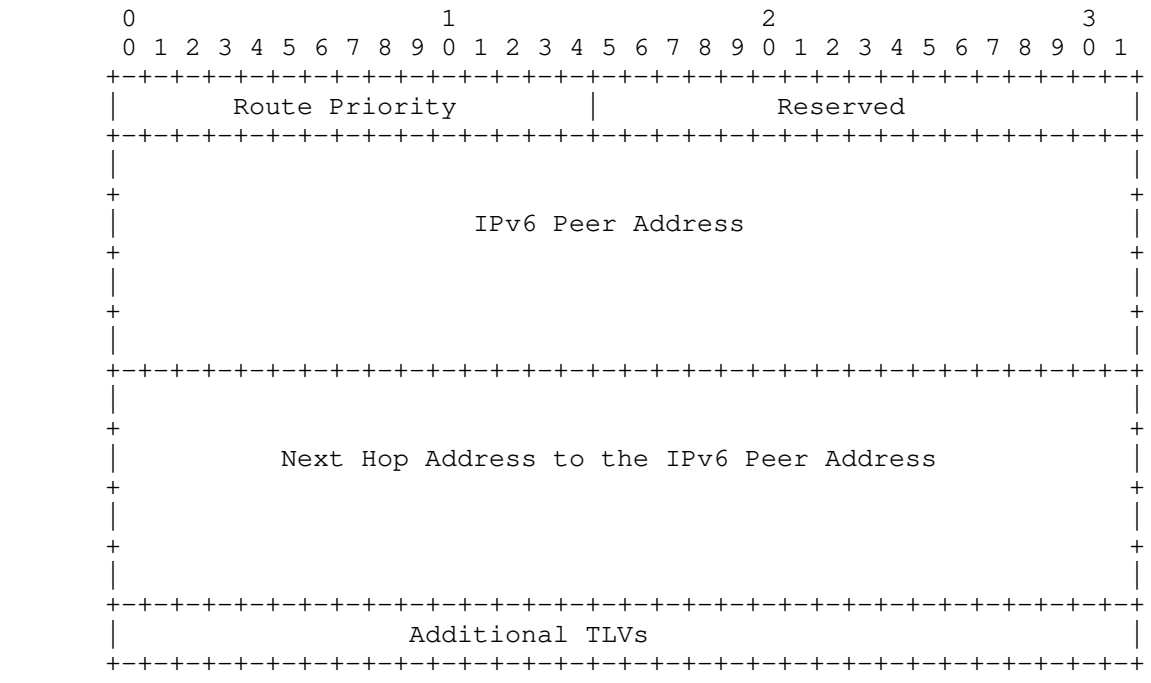


Figure 9: Explicit Peer Route Object Body Format for IPv6

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

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

Peer Address: To indicate the peer address.

Next Hop Address to the Peer: To indicate the next hop address to the corresponding peer.

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

7.4. Peer Prefix Advertisement Object

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

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

Peer Prefix Advertisement Object-Class is TBD16

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

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

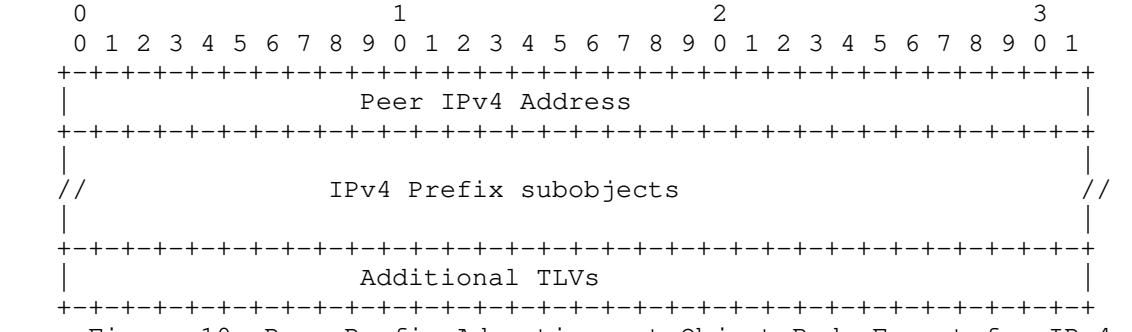


Figure 10: Peer Prefix Advertisement Object Body Format for IPv4

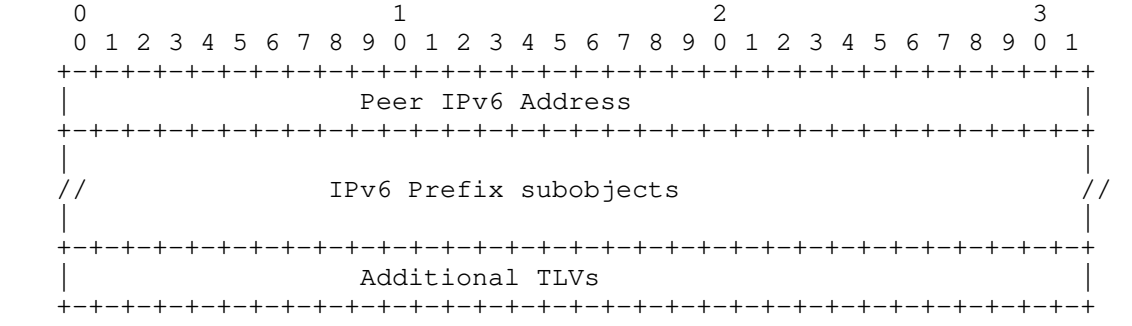


Figure 11: Peer Prefix Advertisement Object Body Format for IPv6

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

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

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

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

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

8. End to End Path Protection

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

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

9. Re-Delegation and Clean up

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

10. BGP Considerations

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

task that appointed by the PCEP protocol and report the status to the PCEP modules.

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

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

11. New Error-Types and Error-Values Defined

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

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

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

12. Deployment Considerations

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

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

When PCE detects one or some of the PCCs are out of control, it should recompute and redeploy the traffic engineering path for native IP on the active PCCs. When PCC detects that it is out of control of the PCE, it should clear the information that initiated by the PCE.

The PCE should assures the avoidance of possible transient loop in such node failure when it deploy the explicit peer route on the PCCs.

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

13. Security Considerations

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

14. IANA Considerations

14.1. Path Setup Type Registry

[RFC8408] created a sub-registry within the "Path Computation Element Protocol (PCEP) Numbers" registry called "PCEP Path Setup Types". IANA is requested to allocate a new code point within this registry, as follows:

Value	Description	Reference
TBD1	Native IP TE Path	This document

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

[I-D.ietf-pce-pcep-extension-for-pce-controller] created a sub-registry within the "Path Computation Element Protocol (PCEP) Numbers" registry to manage the value of the PCECC-CAPABILITY sub-TLV's 32-bits Flag field. IANA is requested to allocate a new bit position within this registry, as follows:

Value	Description	Reference
TBD2(N)	NATIVE-IP-TE-CAPABILITY	This document

14.3. PCEP Object Types

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

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

14.4. PCEP-Error Object

IANA is requested to allocate new error types and error values within the "PCEP-ERROR Object Error Types and Values" sub-registry of the PCEP Numbers registry for the following errors::

Error-Type	Meaning	Error-value	Reference
6	Mandatory Object missing	TBD4:Native IP object missing	This document
10	Reception of an invalid object	TBD3:PCECC NATIVE-IP-TE-CAPABILITY bit is not set	This document
19	Invalid Operation	TBD5:Only one of the BPI,EPR or PPA object can be included in this message	This document
TBD6	Native IP TE failure	TBD7:Peer AS not match TBD8:Peer IP can't be reached TBD9:Local IP is in use TBD10:Remote IP is in use TBD11:Exist BGP session broken TBD12:Explicit Peer Route Error TBD17:EPR/BPI Peer Info mismatch TBD18:BPI/PPA Address Family mismatch TBD19:PPA/BPI Peer Info mismatch	This document

15. Contributor

Dhruv Dhody has contributed the contents of this draft.

16. Acknowledgement

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

17. Normative References

[I-D.ietf-pce-pcep-extension-for-pce-controller]
Li, Z., Peng, S., Negi, M., Zhao, Q., and C. Zhou, "PCEP Procedures and Protocol Extensions for Using PCE as a Central Controller (PCECC) of LSPs", draft-ietf-pce-pcep-extension-for-pce-controller-10 (work in progress), January 2021.

[I-D.ietf-teas-pce-native-ip]
Wang, A., Khasanov, B., Zhao, Q., and H. Chen, "Path Computation Element (PCE) based Traffic Engineering (TE) in Native IP Networks", draft-ietf-teas-pce-native-ip-16 (work in progress), January 2021.

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

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

[RFC4271] Rekhter, Y., Ed., Li, T., Ed., and S. Hares, Ed., "A Border Gateway Protocol 4 (BGP-4)", RFC 4271, DOI 10.17487/RFC4271, January 2006, <https://www.rfc-editor.org/info/rfc4271>.

[RFC4272] Murphy, S., "BGP Security Vulnerabilities Analysis", RFC 4272, DOI 10.17487/RFC4272, January 2006, <https://www.rfc-editor.org/info/rfc4272>.

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

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

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

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

[RFC8253] Lopez, D., Gonzalez de Dios, O., Wu, Q., and D. Dhody, "PCEPS: Usage of TLS to Provide a Secure Transport for the Path Computation Element Communication Protocol (PCEP)", RFC 8253, DOI 10.17487/RFC8253, October 2017, <https://www.rfc-editor.org/info/rfc8253>.

[RFC8281] Crabbe, E., Minei, I., Sivabalan, S., and R. Varga, "Path Computation Element Communication Protocol (PCEP) Extensions for PCE-Initiated LSP Setup in a Stateful PCE Model", RFC 8281, DOI 10.17487/RFC8281, December 2017, <https://www.rfc-editor.org/info/rfc8281>.

[RFC8283] Farrel, A., Ed., Zhao, Q., Ed., Li, Z., and C. Zhou, "An Architecture for Use of PCE and the PCE Communication Protocol (PCEP) in a Network with Central Control", RFC 8283, DOI 10.17487/RFC8283, December 2017, <https://www.rfc-editor.org/info/rfc8283>.

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

[RFC8697] Minei, I., Crabbe, E., Sivabalan, S., Ananthakrishnan, H., Dhody, D., and Y. Tanaka, "Path Computation Element Communication Protocol (PCEP) Extensions for Establishing Relationships between Sets of Label Switched Paths (LSPs)", RFC 8697, DOI 10.17487/RFC8697, January 2020, <https://www.rfc-editor.org/info/rfc8697>.

[RFC8735] Wang, A., Huang, X., Kou, C., Li, Z., and P. Mi, "Scenarios and Simulation Results of PCE in a Native IP Network", RFC 8735, DOI 10.17487/RFC8735, February 2020, <https://www.rfc-editor.org/info/rfc8735>.

Authors' Addresses

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

Email: wangaj3@chinatelecom.cn

Boris Khasanov
Yandex LLC
Ulitsa Lva Tolstogo 16
Moscow
Russia

Email: bhassanov@yahoo.com

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

Email: fsheng@huawei.com

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

Email: tanren@huawei.com

Chun Zhu
ZTE Corporation
50 Software Avenue, Yuhua District
Nanjing, Jiangsu 210012
China

Email: zhu.chun1@zte.com.cn

PCE Working Group
Internet-Draft
Intended status: Standards Track
Expires: August 26, 2021

M. Koldychev
Cisco Systems, Inc.
S. Sivabalan
Ciena Corporation
C. Barth
Juniper Networks, Inc.
S. Peng
Huawei Technologies
H. Bidgoli
Nokia

February 22, 2021

PCEP extension to support Segment Routing Policy Candidate Paths
draft-ietf-pce-segment-routing-policy-cp-03

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

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 26, 2021.

Table of Contents

1. Introduction

Path Computation Element (PCE) Communication Protocol (PCEP) [RFC5440] enables the communication between a Path Computation Client (PCC) and a Path 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 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.

[Editor's Note- Currently it is assumed that each candidate path has only one ERO (SID-List) within the scope of this document. Another document will deal with a way to allow multiple ERO/SID-Lists for a candidate path within PCEP.]

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 optional TLVs based on the association types, that provides 'information' related to the association type.

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 new Association Type (SR Policy Association) and the new TLVs for the ASSOCIATION object, defined in this document, allow a PCEP peer to exchange additional parameters of SR candidate paths and of their associated SR policy. For the SR policy, the parameters are: color and endpoint. For the candidate path, the parameters are: protocol origin, originator, discriminator and preference. [I-D.ietf-spring-segment-routing-policy] describes the concept of SR Policy and these parameters.

The motivation for signaling these parameters is summarized in the following subsections.

3.1. Group Candidate Paths belonging to the same SR policy

Since each candidate path of an SR policy appears as a different LSP (identified via a PLSP-ID) in PCEP, it is useful to group together all the candidate paths that belong to the same SR policy. Furthermore, it is useful for the PCE to have knowledge of the SR candidate path parameters such as color, protocol origin, discriminator, and preference.

3.2. Instantiation of SR policy candidate paths

A PCE needs to instantiate one or more candidate paths on the PCC, as specified in [RFC8281]. Each 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 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, thus saving computation resources on the PCE.

3.4. Minimal signaling overhead

When an SR policy contains multiple 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 candidate path correspond to a unique LSP (identified via PLSP-ID).

For example, if an SR policy has 4 candidate paths, then if the PCE wants to update one of those candidate paths, 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 Candidate Paths and PCEP Associations naturally map to SR Policies. Definition of these mappings is the central purpose of this document.

The mapping between PCEP Associations and SR Policies is always one-to-one. However, the mapping between PCEP Tunnels and SR Candidate Paths may be either one-to-one, or many-to-one. The mapping is one-to-one when the SR Candidate Path has only a single constraint and optimization objective. The mapping is many-to-one when the SR Candidate Path has multiple constraints and optimization objectives. For more details on multiple optimization objectives and constraints, see Section 4.3.

[Editor's Note - Segment-lists within a candidate path are not represented by different PCEP Tunnels. The subject of encoding multiple segment lists within a candidate path is left to another document and is not specified in this document. It is not a good idea to have each segment-list correspond to a different Tunnel, because when the PCC wants to get a path, it must know in advance how many multipaths (i.e., segment-lists) there will be and create that many Tunnels. For example, if the PCC supports 32 multipaths, then it must delegate 32 Tunnels for every candidate path, which may not be scalable.]

A new Association Type is defined in this document, based on the generic ASSOCIATION object. Association type = TBD1 "SR Policy Association" for SR Policy Association Group (SRPAG). The SRPAG Association MUST NOT be used for LSPs that are not part of an SR Policy.

An Association object of SRPAG group contains TLVs that carry Association Information. The association information can be subdivided into three parts: Policy identifiers, Candidate path identifiers, and Candidate path attributes.

Policy Identifiers uniquely identify the SR policy to which a given LSP belongs, within the context of the head-end. Policy Identifiers MUST be the same for all candidate paths in the same SRPAG. Policy Identifiers MUST NOT change for a given LSP during its lifetime. Policy Identifiers MUST be different for different SRPAG associations. When these rules are not satisfied, the PCE MUST send a PCErr message with Error Code = 26 "Association Error", Error Type = TBD6 "Inconsistent SRPAG Identifiers". Policy Identifiers consist of:

- Color of SR policy.
- Endpoint of SR policy.
- Optionally, the policy name.

Candidate Path Identifiers uniquely identify the SR candidate path within the context of an SR policy. Candidate path Identifiers MUST NOT change for a given LSP during its lifetime. 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 Code = 26 "Association Error", Error Type = TBD6 "Inconsistent SRPAG Identifiers". Candidate path Identifiers consist of:

- Protocol Origin of candidate path.
- Originator of candidate path.
- Discriminator of candidate path.
- Optionally, the candidate path name.

Candidate Path Attributes MUST NOT be used to identify the candidate path. Candidate path attributes carry additional information about the candidate path and MAY change during the lifetime of the LSP. Candidate path Attributes consist of:

- Preference of candidate path.

As per the processing rules specified in section 5.4 of [RFC8697], if a PCEP speaker does not support the SRPAG association type, it MUST return a PCErr message with Error-Type 26 "Association Error" and Error-Value 1 "Association-type is not supported". Please note that the corresponding PCEP session is not reset.

4.2. Choice of Association Parameters

The Association Parameters (see Section 2) uniquely identify the Association. In this section, we describe how these are to be set.

The Association Source MUST be set to the headend value of the SR Policy, as defined in [I-D.ietf-spring-segment-routing-policy]. This applies for both PCC-initiated and PCE-initiated candidate paths. The reasoning for this is that if different PCEs could set their own Association Source, then the candidate paths instantiated by different PCEs would by definition be in different PCEP Associations, which contradicts our requirement that the SR Policy is represented by an Association.

If the PCC receives a PCInit message for a non-existing 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:

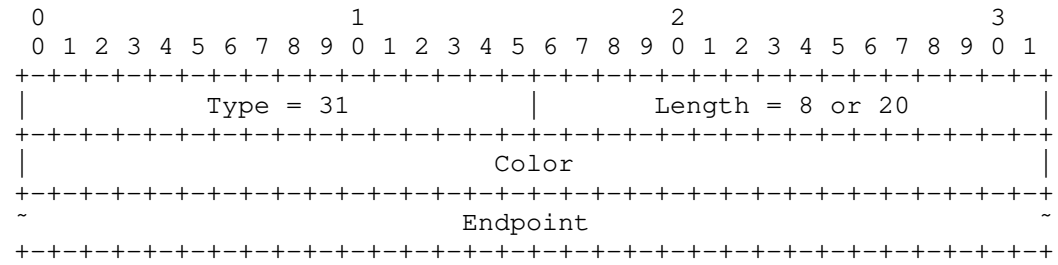


Figure 1: Extended Association ID TLV format

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 head-end.

If the PCEP speaker receives an ASSOCIATION object whose ID and Extended ID do not follow the above specification, then the PCEP speaker MUST send PCErr message with Error-Type 26 "Association Error" and Error-Value 7 "Cannot join the association group".

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.

4.3. Multiple Optimization Objectives and Constraints

In certain scenarios, it is desired for each SR 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 Candidate Paths. This means that multiple PCEP Tunnels are allocated for each SR Candidate Path. Each PCEP Tunnel has its own optimization objective and constraints. When a single SR 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.

5. SR Policy Association Group

Two ASSOCIATION object types for IPv4 and IPv6 are defined in [RFC8697]. The ASSOCIATION object includes "Association type" indicating the type of the association group. This document adds a new Association type.

Association type = TBD1 "SR Policy Association" for SR Policy Association Group (SRPAG).

This Association type is dynamic in nature and created by the PCC or PCE for the candidate paths belonging to the same SR policy (as described in [I-D.ietf-spring-segment-routing-policy]). These associations are conveyed via PCEP messages to the PCEP peer. Operator-configured Association Range MUST NOT be set for this Association type and MUST be ignored.

SRPAG MUST carry additional TLVs to communicate Association Information. This document specifies four new TLVs to carry Association Information: SRPOLICY-POL-NAME TLV, SRPOLICY-CPATH-ID

TLV, SRPOLICY-CPATH-NAME TLV, SRPOLICY-CPATH-PREFERENCE TLV. These four TLVs encode the SR Policy Name, Candidate Path Identifiers, Candidate Path Name, and Candidate Path Preference, respectively. Of these new TLVs, only SRPOLICY-CPATH-ID TLV is mandatory. When the mandatory TLV is missing from the SRPAG association object, the PCE MUST send a PCErr message with Error Code = 26 "Association Error", Error Type = TBD7 "Missing mandatory SRPAG TLV".

A given LSP MUST belong to at most one SRPAG, since a candidate path cannot belong to multiple SR policies. If a PCEP speaker receives a PCEP message with more than one SRPAG for an LSP, then the PCEP speaker MUST send a PCErr message with Error-Type 26 "Association Error" and Error-Value TBD8 "Same LSP in multiple SRPAG". If the message is a PCRpt message, then the PCEP speaker MUST close the PCEP connection. Closing the PCEP connection is necessary because ignoring PCRpt messages may lead to inconsistent LSP DB state between the two PCEP peers.

If the PCEP speaker receives the SRPAG association when the SR capability (as per [RFC8664] or [I-D.ietf-pce-segment-routing-ipv6]) was not exchanged, the PCEP speaker MUST send a PCErr message with Error-Type 26 "Association Error" and Error-Value TBD9 "Use of SRPAG without SR capability exchange". If the Path Setup Type (PST) of the LSP in SRPAG is not set to SR or SRv6, then the PCEP speaker MUST send a PCErr message with Error-Type 26 "Association Error" and Error-Value TBD10 "non-SR LSP in SRPAG".

5.1. SR Policy Name TLV

The SRPOLICY-POL-NAME TLV is an optional TLV for the SRPAG 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.

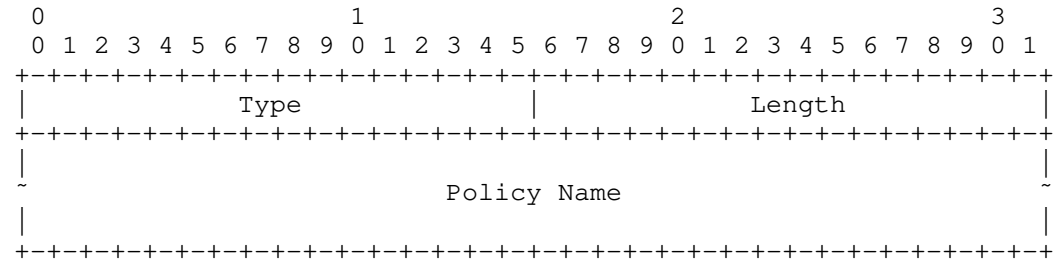


Figure 2: The SRPOLICY-POL-NAME TLV format

Type: TBD2 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.

Policy Name: 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. SR Policy Candidate Path Identifiers TLV

The SRPOLICY-CPATH-ID TLV is a mandatory TLV for the SRPAG 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.

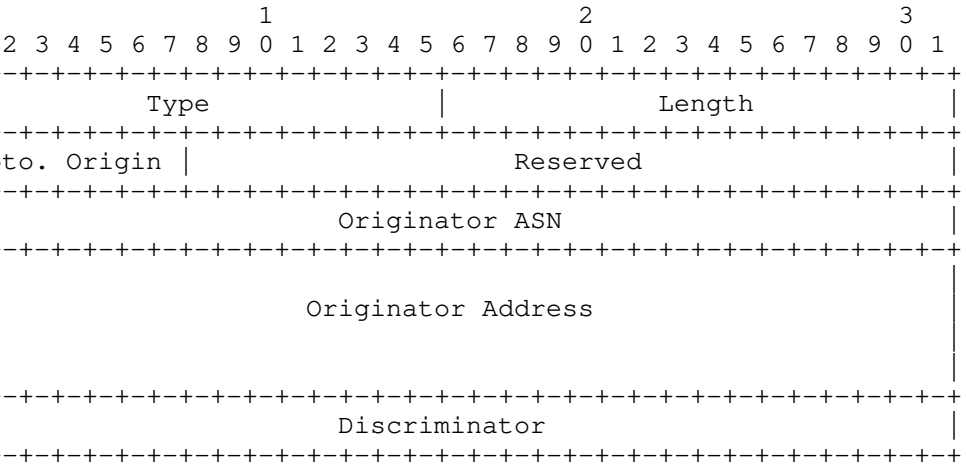


Figure 3: The SRPOLICY-CPATH-ID TLV format

Type: TBD3 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.

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

Originator ASN: Represented as 4 byte number, part of the originator identifier, as specified in [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.

5.3. SR Policy Candidate Path Name TLV

The SRPOLICY-CPATH-NAME TLV is an optional TLV for the SRPAG 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.

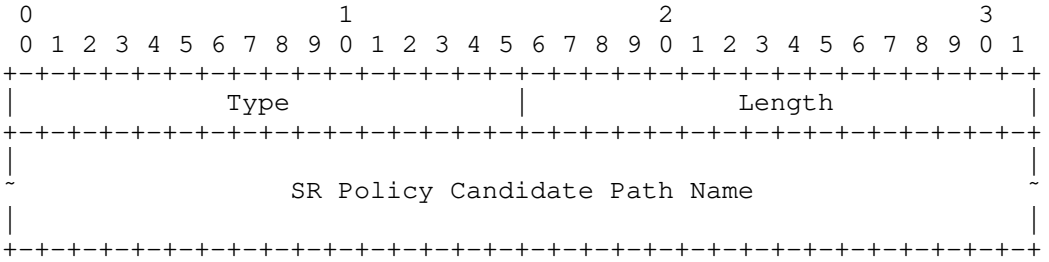


Figure 4: The SRPOLICY-CPATH-NAME TLV format

Type: TBD4 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.

5.4. SR Policy Candidate Path Preference TLV

The SRPOLICY-CPATH-PREFERENCE TLV is an optional TLV for the SRPAG 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.

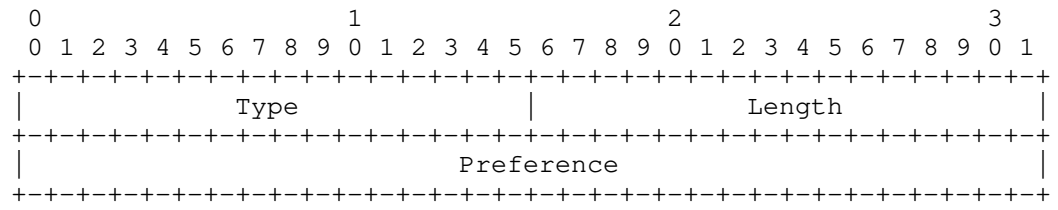


Figure 5: The SRPOLICY-CPATH-PREFERENCE TLV format

Type: TBD5 for "SRPOLICY-CPATH-PREFERENCE" TLV.

Length: 4.

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

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

6. Examples

6.1. 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 SRPAG ASSOCIATION object and TLVs in the PCReq message.

2. PCE returns the path in PCRep message, and echoes back the SRPAG object that was used in the computation.

PCRpt and PCUpd messages are exchanged in the following sequence:

1. PCC sends PCRpt message to the PCE, including the LSP object and the SRPAG ASSOCIATION object.

2. PCE computes path, possibly making use of the Association Information from the SRPAG ASSOCIATION object.

3. PCE updates the SR policy candidate path's ERO using PCUpd message.

6.2. 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 SRPAG ASSOCIATION object 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.

6.3. PCE Initiated SR Policy with single candidate-path

A candidate-path is created using the following steps:

1. PCE sends PCInitiate message, containing the SRPAG Association object. The Association Source is set to the IP address of the PCC and the Association ID is set to 0, as described in Section 4.2.

2. PCC uses the color, endpoint and preference from the SRPAG object 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 SRPAG Association object. 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.

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.

6.4. 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 SRPAG Association object is sent for every LSP in the PCInitiate message. The Association Source is set to the IP address of the PCC and the Association ID is set to 0, as described in Section 4.2.

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 SRPAG Association object. 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.

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.

7. IANA Considerations

7.1. 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
TBD1	SR Policy Association	This.I-D

7.2. 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
TBD2	SRPOLICY-POL-NAME	This.I-D
TBD3	SRPOLICY-CPATH-ID	This.I-D
TBD4	SRPOLICY-CPATH-NAME	This.I-D
TBD5	SRPOLICY-CPATH-PREFERENCE	This.I-D

7.3. PCEP Errors

This document defines five 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
26	Association Error		[RFC8697]
		TBD6: Inconsistent SRPAG Identifiers	This.I-D
		TBD7: Missing mandatory SRPAG TLV	This.I-D
		TBD8: Same LSP in multiple SRPAG	This.I-D
		TBD9: Use of SRPAG without SR capability exchange	This.I-D
		TBD10: non-SR LSP in SRPAG	This.I-D

8. 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".

8.1. Cisco

- Organization: Cisco Systems
- Implementation: Head-end and controller.
- Description: An experimental code-point is currently used.
- Maturity Level: Proof of concept.
- Coverage: Full.
- Contact: mkoldych@cisco.com

9. 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 SRPAG Association object, 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 SRPAG association could provides 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.

10. Acknowledgement

Would like to thank Stephane Litkowski, Praveen Kumar and Tom Petch for review comments.

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>.

[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", draft-ietf-spring-segment-routing-policy-09 (work in progress), November 2020.

[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., Negi, M., Achaval, D., and H. Kotni, "PCEP Operational Clarification", draft-koldychev-pce-operational-02 (work in progress), August 2020.

11.2. Informative References

[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>.

[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", draft-ietf-pce-segment-routing-ipv6-08 (work in progress), November 2020.

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

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: September 9, 2021

M. Koldychev
Cisco Systems, Inc.
S. Sivabalan
Ciena Corporation
C. Barth
Juniper Networks, Inc.
S. Peng
Huawei Technologies
H. Bidgoli
Nokia

March 8, 2021

PCEP extension to support Segment Routing Policy Candidate Paths
draft-ietf-pce-segment-routing-policy-cp-04

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

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 9, 2021.

Table of Contents

1. Introduction

Path Computation Element (PCE) Communication Protocol (PCEP) [RFC5440] enables the communication between a Path Computation Client (PCC) and a Path 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 TBD1, 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".

An SRPAT ASSOCIATION carries three pieces of information: SR Policy Identifiers, SR Policy Candidate Path Identifiers, and SR Policy Candidate Path Attributes.

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.

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.

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.

As per the processing rules specified in section 5.4 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".

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 = TBD1 "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 TBD1 "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:

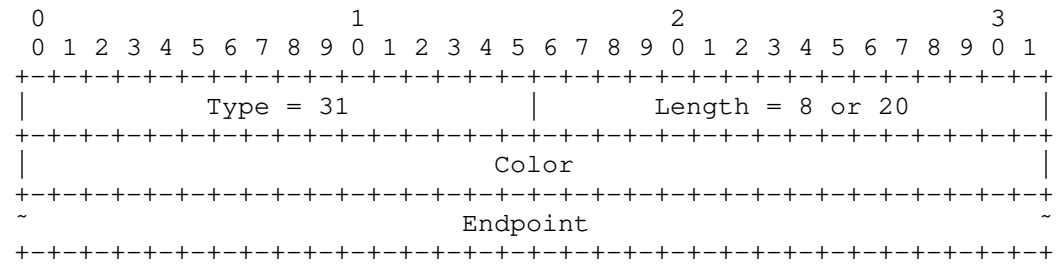


Figure 1: Extended Association ID TLV format

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.

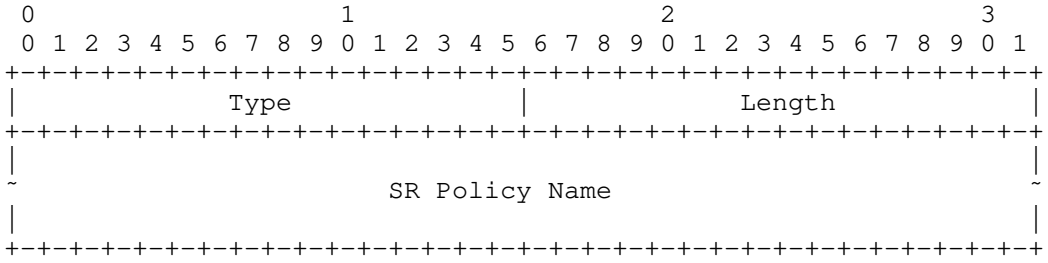


Figure 2: The SRPOLICY-POL-NAME TLV format

Type: TBD2 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.

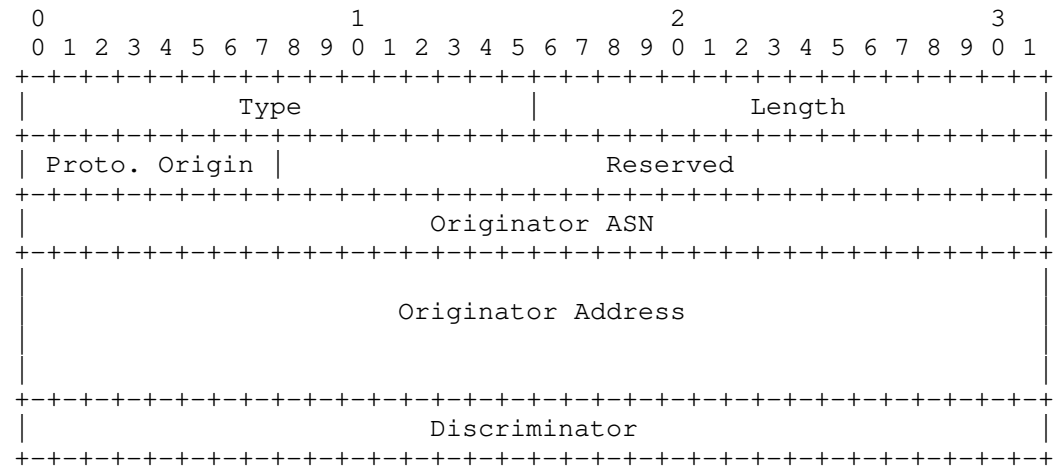


Figure 3: The SRPOLICY-CPATH-ID TLV format

Type: TBD3 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.

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

Originator ASN: Represented as 4 byte number, part of the originator identifier, as specified in [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.

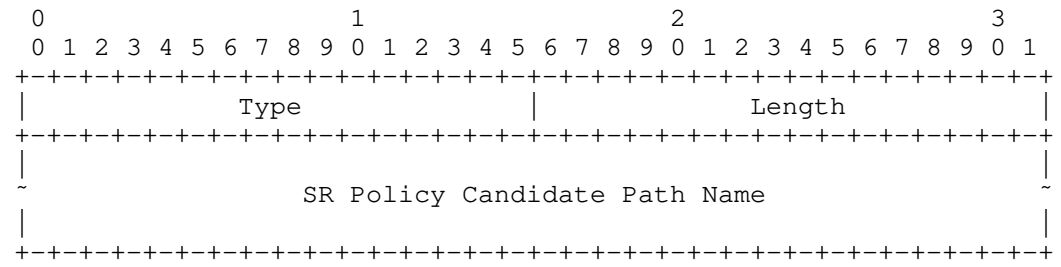


Figure 4: The SRPOLICY-CPATH-NAME TLV format

Type: TBD4 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.

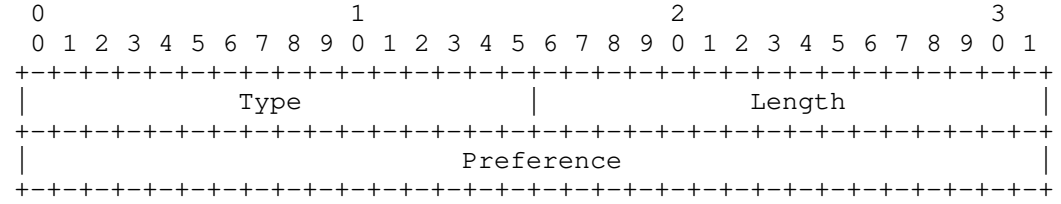


Figure 5: The SRPOLICY-CPATH-PREFERENCE TLV format

Type: TBD5 for "SRPOLICY-CPATH-PREFERENCE" TLV.

Length: 4.

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

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

6. Examples

6.1. 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.

6.2. 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.

6.3. 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.

6.4. 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.

7. IANA Considerations

7.1. 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
TBD1	SR Policy Association	This.I-D

7.2. 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
TBD2	SRPOLICY-POL-NAME	This.I-D
TBD3	SRPOLICY-CPATH-ID	This.I-D
TBD4	SRPOLICY-CPATH-NAME	This.I-D
TBD5	SRPOLICY-CPATH-PREFERENCE	This.I-D

7.3. 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 Identifers Mismatch	This.I-D
		TBD8: SR Policy Candidate Path Identifiers Mismatch	This.I-D

8. 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".

8.1. 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

8.2. 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

9. 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.

10. Acknowledgement

Would like to thank Stephane Litkowski, Praveen Kumar and Tom Petch for review comments.

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>.

[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", draft-ietf-spring-segment-routing-policy-09 (work in progress), November 2020.

[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., Negi, M., Achaval, D., and H. Kotni, "PCEP Operational Clarification", draft-koldychev-pce-operational-02 (work in progress), August 2020.

[I-D.koldychev-pce-multipath]
Koldychev, M., Sivabalan, S., Saad, T., Beeram, V., Bidgoli, H., Yadav, B., and S. Peng, "PCEP Extensions for Signaling Multipath Information", draft-koldychev-pce-multipath-04 (work in progress), October 2020.

11.2. Informative References

[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>.

[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", draft-ietf-pce-segment-routing-ipv6-08 (work in progress), November 2020.

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

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

Path Computation Element Working Group
Internet-Draft
Intended status: Standards Track
Expires: August 26, 2021

O. Dugeon
J. Meuric
Orange Labs
Y. Lee
Samsung Electronics
D. Ceccarelli
Ericsson

February 22, 2021

PCEP Extension for Stateful Inter-Domain Tunnels
draft-ietf-pce-stateful-interdomain-01

Abstract

This document specifies how to use a Backward Recursive or Hierarchical method to derive inter-domain paths in the context of stateful Path Computation Element (PCE). The mechanism relies on the PCInitiate message to set up independent paths per domain. Combining these different paths together enables them to be operated as end-to-end inter-domain paths, without the need for a signaling session between inter-domain border routers. For this purpose, this document defines a new Stitching Label, new Path Setup Types, new Association Type, and a new PCEP communication Protocol (PCEP) Capability.

Table of Contents

1. Introduction

The PCE working group has produced a set of RFCs to standardize the behavior of the Path Computation Element ([RFC4655] and [RFC5440]) as a tool to help MultiProtocol Label Switching - Traffic Engineering (MPLS-TE)/Generalized MPLS (GMPLS) Label Switched Paths (LSPs) and Segment Routing paths placement. This also includes the ability to compute inter-domain LSPs or Segment Routing paths following a distributed BRPC [RFC5441] or hierarchical H-PCE [RFC6805] approach. Such inter-domain paths could then serve as an Explicit Route Object (ERO) input for the RSVP-TE signaling to set up the tunnels within the underlying network. Three kinds of inter-domain paths could be established:

- Contiguous tunnel ([RFC3209] and [RFC3473]): The RSVP-TE signaling crosses the boundary between two domains. This kind of tunnel is not recommended mostly for security and scalability purpose. In addition, the initiating domain imposes huge constraints on subsequent domains, because they undergo the tunnel request without being able to control it.

- Stitching tunnel ([RFC5150]): Each domain establishes in its own network the corresponding part of the inter-domain path independently. Then, a second end-to-end RSVP-TE Path message is sent by the initiating domain to stitch the different tunnel parts to form the inter-domain path.

- Nesting tunnel ([RFC4206]): This is similar to the stitching mode but, this time, with the possibility to set up tunnel hierarchy.

However, these inter-domain paths depend on signaling using RSVP-TE to be set up, but it is not common to allow signaling across administrative domain borders, especially in operational networks.

For Segment Routing, issues are different as there is no signaling between routers. First, a segment path depends on a stack of segment identifiers but, in an inter-domain path, this stack may become too large with respect to hardware constraint. If Extensions for Segment Routing [RFC8664] takes into account the Maximum Stack Depth (MSD), a PCE may be unable to find a solution when it computes an end-to-end inter-domain path. The second issue is related to the path confidentiality because all Node-SID must be stacked by the head end router while some of the Node-SIDs are associated to routers of the next domains. It is clear that operators would not disclose details

of their network, which includes Node-SIDs. Thus, it is not possible to stack remote labels for an end-to-end inter-domain path even if MSD constraint is respected.

The purpose of this document is to take the benefit of Active Stateful PCE [RFC8231] and PCE-Initiated [RFC8281] modes to stitch or nest inter-domain paths directly using PCEP between domains' PCEs while avoiding the use of another signaling between inter-domain border nodes. The mechanism keeps each operator free to independently set up their respective part of the inter-domain paths, i.e. the signaling (for MPLS-TE and GMPLS) is scoped on a per domain basis, individually.

The PCInitiate message is used from destination domain to source domain, to recursively set up the end-to-end tunnel. PCRep message is used to convey the specific labels or SIDs to automatically stitch or nest the different local LSPs. And PCRep in conjunction with PCUpd messages are used to report, maintain, modify and tear down inter-domain paths. This method is also applicable to Segment Routing to build inter-domain segment paths. To enable this mechanism, this document defines a new Stitching Label, new Path Setup Types, new Association Type, and a new PCEP communication Protocol (PCEP) Capability.

<Editor's note: the replacing encoding to use instead of PST is to be discussed with the WG.>

1.1. General Assumptions

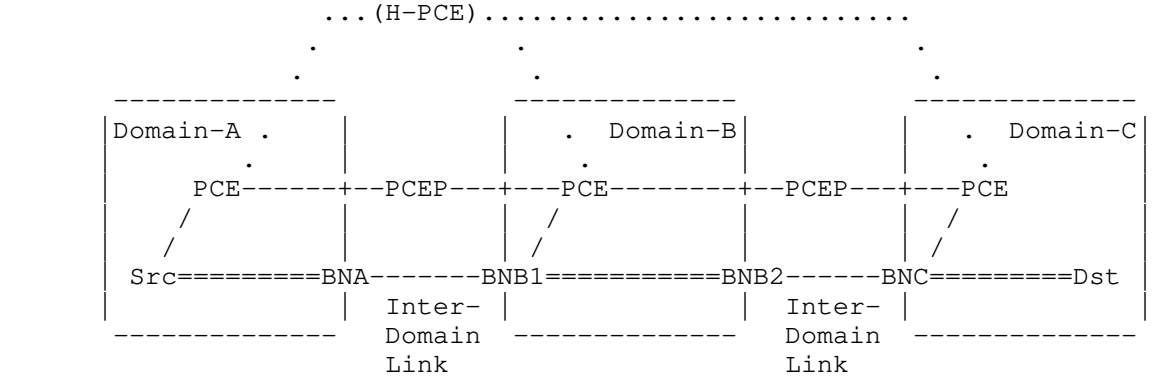
In the remainder of this document, the same references as per BRPC [RFC5441] are used and the following set of assumptions are made (see figure below):

- Domain refers to administrative partitions, i.e. an IGP area or an Autonomous System (AS).
- Inter-domain path is used to refer to a path that crosses two or more different domains as defined previously,
- At least one PCE is deployed in each domain. These PCEs are all active stateful-capable and can request to enforce LSPs in their respective domain by means of PCInitiate messages.
- LSRs, including border nodes, are PCC-enabled and support active stateful mode. PCEP sessions are established between these routers and their domains' PCE.

- Each PCE establishes a PCEP session with its respective neighbor domains' PCEs. The way a PCE discovers its neighboring PCEs is out of the scope of this document.

- PCEs are able to compute an end-to-end path as per BRPC procedure [RFC5441] or as per H-PCE procedure (stateless [RFC6805] or stateful [RFC8751]).

- "Path" is a generic term to refer to both LSP setup by mean of RSVP-TE or Segment Path in a Segment Routing network.



Example of the representation of 3 domains with 3 PCEs

Operations, according to the figure above, are as follow:

1. The PCEs in Domain-A, Domain-B, and Domain-C communicate using PCEP either directly, as shown, using BRPC or with a parent PCE if using H-PCE.

2. The PCE in Domain-A selects an end-to-end domain path. It tells the PCE in Domain-B that the path will be used, and that PCE passes the information on to the PCE in Domain-C.

3. Each of the PCEs use PCEP to instruct the segment head ends backward from destination to source:

   A. In Domain-C, the PCE instructs the ingress Border Node, BNC, with the path to reach the Destination. The instructions also ask BNC to provide the incoming label or SID that will be stitched to the intra-domain path. Once done, PCE reports this label or SID to PCE of Domain-B.

   B. In Domain-B, the PCE instructs the ingress Border Node, BNB1, with the path to reach the egress Border Node, BNB2. The

instructions also tell BNB1 the label or SID to use on the inter-domain link to BNC and ask to provide the incoming label or SID that will be stitched to the intra-domain path. Once done, PCE reports this label or SID to PCE of Domain-A.

C. In Domain-A, the PCE instructs the Source node with the path to use to reach Border Node, BNA. The instructions also include the label or SID to use on the inter-domain link to BNB1.

1.2. Terminology

ABR: Area Border Routers. Routers used to connect two IGP areas (areas in OSPF or levels in IS-IS).

AS: Autonomous System

ASBR: Autonomous System Border Router. Router used to connect together ASes (of the same or different service providers) via one or more inter-AS links.

Border Node (BN): a boundary node is either an ABR in the context of inter-area TE or an ASBR in the context of inter-AS TE.

BN-en(i): Entry BN of domain(i) connecting domain(i-1) to domain(i) along a determined sequence of domains. Multiple entry BN-en(i) could be used to connect domain(i-1) to domain(i).

BN-ex(i): Exit BN of domain(i) connecting domain(i) to domain(i+1) along a determined sequence of domains. Multiple exit BN-ex(i) could be used to connect domain(i) to domain(i+1).

Domains: Autonomous System (AS) or IGP Area. An Autonomous System is composed by one or more IGP area.

ERO(i): The Explicit Route Object scoped to domain(i)

IGP-TE: Interior Gateway Protocol with Traffic Engineering support. Both OSPF-TE and IS-IS-TE are identified in this category.

Inter-domain path: A path that crosses two or more domains through a pair of Border Node (BN-ex, BN-en).

LK(i): A Link that connect BN-ex(i-1) to BN-en(i). Note that BN-ex(i-1) could be connected to BN-en(i) by more than one link. LK(i) identifies which of the multiple links will be used for the inter-domain path setup. For inter-AS scenario, LK(i) represents the link between ASBR of domain i to the ASBR of domain i-1. For inter-area

scenario, LK(i) is present only in IS-IS networks and represents the link between ABR of region L1, reciprocally L2, to the ABR of region L2, reciprocally L1.

Local path: A path that does not cross a domain border. It is set up either from entry BN-en, to output BN-ex or between both. This path could be enforce by means of RSVP-TE signaling or Segment Routing labels stack.

Local path(i): A Local path of domain(i)

PLSP-ID(i): A PLSP-ID that identifies, in the domain(i), the local part of an inter-domain path.

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.

PCE(i) is a PCE within the scope of domain(i).

PST: Path Setup Type

R(i,j): The router j of domain i

Stitching Label (SL): A dedicated label that is used to stitch two RSVP-TE LSPs or two Segment Routing paths.

SL(i): A Stitching Label that links domain(i-1) to domain(i).

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. Stitching Label

This section introduces the concept of Stitching Label that allows stitching and nesting of local paths in order to form an inter-domain path that cross several different domains.

2.1. Definition

The operation of stitch or nest a local path(i) to a local path(i+1) in order to form and inter-domain path mainly consists in defining the label that the output BN-ex(i) will use to send its traffic to the entry BN-en(i+1). Indeed, the entry BN-en(i+1) needs to identify the incoming traffic (e.g. IP packets), in order to know if this traffic must follow the local path(i+1) or not. Forwarding Equivalent Class (FEC) could be used for that purpose. But, when

stitching or nesting tunnels, the FEC is reduced to the incoming label that the entry BN-en(i+1) has chosen for the local path(i+1).

In this document, we introduce the term of "Stitching Label (SL)" to refer to this label. Such label is usually exchanged between output BN-ex(i) and entry BN-en(i+1) with the RSVP-TE signaling. But, as we want to avoid to use RSVP-TE signaling due to operational constraints, and allow compatibility support for Segment Routing, this Stitching Label is here conveyed by PCEP. In fact, the Explicit Route Object (ERO) and the Record Route Object (RRO) are already defined in order to transport (G)MPLS labels (for RSVP-TE or Segment Routing) in the PCEP signaling. Thus, the Stitching Label could be conveyed in the ERO and RRO without any modification of PCEP nor PCEP Objects.

As per RFC4003 [RFC4003], the Stitching Label will be conveyed as a companion of a link identifier (e.g. an IP address for numbered links). In our case, this is one of the endpoint IDs of the link LK(i) which connects BN-ex(i) to BN-en(i+1) and carries the traffic from the domain(i) to domain(i+1). It is left to implementation to select which of the two endpoint IDs of the link LK(i) is used.

2.2. Inter-domain LSP-TYPE

Even if PCEP could convey the Stitching Label, a PCC is not aware that a PCE requests or provides such a label. For that purpose, this specification relies on the use of the PST as defined in [RFC8408] with new values (See IANA section of this document) defined as follow:

- TBD1: Inter-Domain TE end-to-end path is set up using Backward Recursive or Hierarchical method. This new PST value MUST be set in a PCInitiate messages sends by a PCE(i-1) to its neighbor PCE(i) in the Backward Recursive method or by the Parent PCE to the Child PCE(i) to initiate a new inter-domain path. In its response, the neighbor PCE(1) or Child PCE(i) MUST return a Stitching Label SL with an identifier of the associated link in the RRO of the PCRpt message to PCE(i-1) or Parent PCE.

- TBD2: Inter-Domain TE local path is set up using RSVP-TE. This new PST value MUST be set in the PCInitiate message sends by a PCE(i) requesting to a PCC of domain(i) to initiate a new local path(i) which is part of an inter-domain path. This PST value MUST be used by the PCE(i) only after receiving a PCInitiate message with an PST equal to TBD1 from a neighbor PCE(i-1) in the Backward Recursive method or Parent PCE in the Hierarchical method. In its response, the PCC of domain(i) MUST return a

Stitching Label SL with the an identifier of associated link in the RRO of the PCRpt message.

- TBD3: Inter-Domain TE local path is set up using Segment Routing. This new PST value MUST be set in the PCInitiate message sends by a PCE(i) requesting to a PCC of domain(i) to initiate a new Segment Routing path which is part of and inter-domain Segment Routing path. This PST value MUST be used by the PCE(i) only after receiving a PCInitiate message with an PST equal to TBD1 from a neighbor PCE(i-1). In its response, the PCC MUST return a Stitching Label SL with an identifier of the associated link in the RRO of the PCRpt message.

<Editor's note: the replacing encoding to use instead of PST is to be discussed with the WG.>

3. Backward Recursive PCInitiate Procedure

This section describes how to set up inter-domain paths that cross different domains by using a Backward Recursive method. It is compatible with the inter-domain path computation by means of the BRPC procedure as describe in RFC5441 [RFC5441].

3.1. Mode of Operation

This section describes how PCInitiate and PCRpt messages are combined between PCE in order to set up inter-domain paths between a source domain(1) to a destination domain(n). S and D are respectively the source and destination of the inter-domain path. Domain(1) and domain(n) are different and connected through 0 (i.e. direct connection when n = 2) or more intermediate domains denoted domain(i) with i = [2, n-1].

First, the PCE(1) runs standard BRPC algorithm as per RFC5441 [RFC5441] with its neighbor PCEs in order to compute the inter-domain path from S to D, where S and D are respectively a node in the domain(1) and domain(n). Path Key confidentiality as per RFC5520 [RFC5520] SHOULD be used to obfuscate the detailed ERO(i) of the different domains(i). The resulting ERO is in the form {S, PKS(1), BN-ex(1), ..., BN-en(i), PKS(i), BN-ex(i), ..., BN-en(n), PKS(n), D} when Path Key is used and of the form {S, R(1,1), ..., R(1,k), BN-ex(1), ..., BN-en(i), R(i,1), ..., R(i,l), BN-ex(i), ..., BN-en(n), R(n,1), ..., R(n,m), D} otherwise . As subsequent domains are not aware about the computed end-to-end ERO in case of Virtual Source Path trees (VSPTs), the final ERO selected by the PCE(1) MUST be sent in the PCInitiate message to indicate to the subsequent PCEs which path has been finally chosen. PCE(1) MUST ensure that this ERO is self comprehensive by subsequent PCEs. Indeed, when a PCE(i)

receives the ERO, it MUST be able to verify that this ERO matches its own scope and to determine the PCE(i+1). When Path Key is used, PCEs MUST encode the Path Key with a reachable IP address so that previous PCEs in the AS chain are able to join them. When Path Key is not used, the PCEs MUST be able to retrieve an IP address of the next PCE corresponding to the ERO (e.g., relying on a per prefix table).

The complete procedure with Path Key follows the different steps described below:

Steps 1: Initialization

Once ERO(S, D) is computed, PCE(1) sends a PCInitiate message to PCE(2) containing an ERO equal to {S, PKS(2), ..., PKS(i), ..., PKS(n), D}, PST = TBD1 and End-Points Object = (S, D). The ERO corresponds to the one PCE(1) has received from PCE(2) during the BRPC process in which only Path Key are kept. In case of multiple EROs, i.e. VSPT, PCE(1) has chosen one of them and used the selected one for the PCInitiate message. PKS(i) could be replaced by the full ERO description if Path Key is not used by PCE(i).

When PCE(i) receives a PCInitiate message from domain(i-1) with PST = TBD1 and ERO = {PKS(i), PKS(i+1), ..., PKS(n), D)}, it sends a PCInitiate message to PCE(i+1) with a popped ERO and records its received PKS(i) part. All PCE(i)s generate the appropriate PCInitiate message to PCE(i+1) up to PCE(n), i.e. to the destination domain(n).

Steps 2: Actions taken at the destination domain(n) by PCE(n)

1. When a PCInitiate message reaches the destination domain(n), PCE(n) retrieves the ERO from the PKS(n) if necessary and sends to BN-en(n) a PCInitiate message with the ERO(n) = {BN-en(n), ..., D}, PST = TBD2 and End-Points Object = {BN(n), D} in order to inform the PCC BN-en(n) that this local path(n) is part of an inter-domain path.

2. When the PCC BN-en(n) receives the PCInitiate message from its PCE(n), it sets up the local path from entry BN-en(n) to D by means of RSVP-TE signaling with the given ERO(n).

3. Once the tunnel is set up, BN-en(n) chooses a free label for the Stitching Label SL(n) and adds a new entry in its MPLS L(F)IB with this SL(n) label. Then, it sends a PCRpt message to its PCE(n) with an RRO equal to {[LK(n), SL(n)], RRO(n)} and PLSP-ID(n).

4. Once PCE(n) receives the PCRpt from the PCC BN-en(n) with the RRO, PLSP-ID and PST = TBD2, it sends to the PCE(n-1) a PCRpt containing the RRO equal to {[LK(n), SL(n)]} and PLSP-ID(n). PCE(n) MAY add {PKS(n), D} in the RRO.

Steps i: Actions performed by all intermediate domains(i), for i = 2 to n-1

1. When the PCE(i) receives a PCRpt message from domain(i+1) with PST = TBD1, RRO = {[LK(i+1), SL(i+1)]} and PLSP-ID(i+1), it retrieves the ERO(i) from the PKS(i), recorded in step 1, and sends to the PCC BN-en(i) a PCInitiate message with ERO = {ERO(i), [LK(i+1), SL(i+1)]}, PST = TBD2 and End-Points Object = {BN-en(i), BN-ex(i)} in order to inform the PCC BN-en(i) that this local path(i) is part of an inter-domain path.

2. When the PCC BN-en(i) receives the PCInitiate message from its PCE(i), it sets up the local path from BN-en(i) to BN-ex(i) by means of RSVP-TE signaling with the given ERO(i).

3. Egress Control mechanism, as per RFC4003 section 2.1 [RFC4003], is used to instruct the egress node of domain(i), i.e. BN-ex(i), to forward packets belonging to this tunnel with the Stitching Label. Both the Stitching Label and the identifier of the interface are carried in the ERO = {..., [LK(i+1), SL(i+1)]} as the last SubObject in conformance to [RFC4003]. As a result, BN-ex(i) installs in its MPLS L(F)IB the SWAP instruction to label SL(i+1) with forward to LK(i+1).

4. Once the tunnel is set up, PCC BN-en(i) chooses a free label for the Stitching Label SL(i) and adds a new entry in its MPLS L(F)IB with this SL(i) label. Then, it sends a PCRpt message to its PCE(i) with an RRO equal to {[LK(i), SL(i)], RRO(i)} and PLSP-ID(i).

5. Once PCE(i) receives the PCRpt from the PCC BN-en(i) with the RRO and PST = TBD2, it sends to PCE(i-1) a PCRpt message containing the RRO equal to {[LK(i), SL(i)]} and the PLSP-ID(i). PCE(i) MAY add {PKS(i), ..., PKS(n)} in the RRO.

Steps n: Actions performed at the source domain(1) by PCE(1)

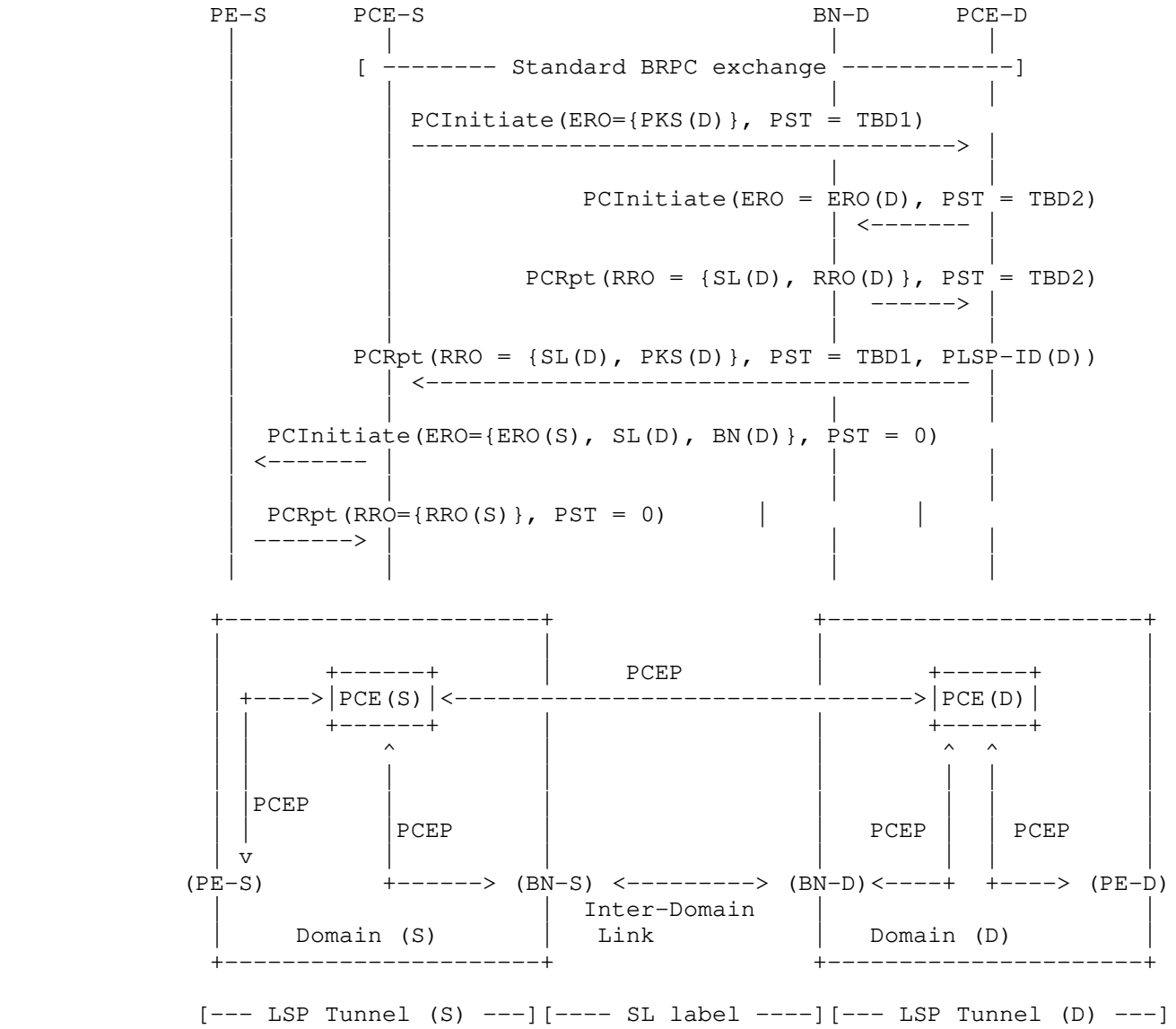
Once PCE(1) receives the PCRpt message from PCE(2) with the RRO containing the label SL(2), it sends a PCInitiate message to PCC node S with ERO equal to {ERO(1), [LK(2), SL(2)]}, PST = 0 and End-Points Object = {S, BN-ex(1)}. This time, the PST is equal to 0 as the PCC S does not need to return a Stitching Label SL, because it is the head-

end of the inter-domain path. A usual PCRpt message is sent back to PCE(1) by the PCC node S.

3.2. Example

In the figure below, two different domains S and D are interconnected through BN respectively BN-S and BN-D. PE-S and PE-D are edge routers. All routers in the figure are connected to their respective PCE through PCEP. In this example, we consider that PCE(S) needs to set up an inter-domain path between PE-S and PE-D acting as source and destination of the path. To simplify the figure, neither intermediate routers between (PE-S, BN-S), (BN-D and PE-D), nor RSVP-TE messages are represented, but they are all presents. The following notation is used (in this example, we use the PKS for the sake of simplicity):

- PKS(D) = Path Key corresponding to the path from BN(D) to PE-D
- ERO(D) = Explicit Route Object corresponding to the path from BN(D) to PE-D, retrieved from PKS(D)
- RRO(D) = Record Route Object of the local path(D) from BN(D) to PE-D
- SL(D) = Stitching Label for the local path from BN(D) to PE-D
- ERO(S) = Explicit Route Object corresponding to the path from PE-S to BN(S)
- RRO(S) = Record Route Object of local path(S) from PE-S to BN(S)



Example of inter-domain path setup between two domains

3.3. Completion Failure of Inter-domain Path Setup Procedure

In case of error during path setup, PCRpt and or PCErr messages MUST be used to signal the problem to the neighbor PCE domain backward. In particular, if the new PST values defined in this document are not

supported by the neighbor PCE or the PCC, the PCE, respectively the PCC, MUST return a PCErr message with Error-Type = 21 (TE path setup error) and Error-Value = 1 (Unsupported path setup type) to its neighbor PCE. If a PCE(i) receives a PCInitiate message from its peer PCE(i-1) without PST set to TBD1 or PST set to a value different from TBD1, it MUST return a PCErr message with Error-Type = 21 (TE path setup error) and Error-Value = 1 (Unsupported path setup type) to its peer PCE(i-1).

Following a PCInitiate message with PST set to TBD1, if a PCC or a PCE returns no RRO, or an RRO without the Stitching Label SL and an identifier of the associated link, the PCE MUST return a PCErr message with Error-Type = 21 (TE path setup error) and Error-Value = TBD5 (Mandatory Stitching Label missing in the RRO).

In case of completion failure, the PCE(i) MUST propagate the PCErr message up to the PCE(1). In turn, PCE(1) MUST send a PCInitate message (R flag set in the SRP Object as per [RFC8281]) to tear down this inter-domain path from its neighbor PCEs. PCE(i) MUST propagate the PCInitiate message and remove its local path by means of PCInitiate message to its PCC BN-en(i) and send back PCRpt message to PCE(i-1).

In case of error in domain(i+1), PCE(i) MAY add the AS number of domain(i+1) in the RRO to identify the faulty domain.

4. Hierarchical PCInitiate Procedure

This section describes how to set up inter-domain paths that cross different domains by using a hierarchical method. It is compatible with inter-domain path computation as described in [RFC6805].

4.1. Mode of Operation

This section describes how PCInitiate and PCRpt messages are combined between PCEs in order to set up inter-domain paths between a source domain(1) to a destination domain(n). S and D are respectively the source and destination of the inter-domain path. Domain(1) and domain(n) are different and connected through 0 or more intermediate domains denoted domain(i) with i = (2, n-1). Domains are directly connected when n = 2.

First, the Parent PCE contacts its Child PCE as per [RFC6805] in order to compute the inter-domain path from S to D, where S and D are respectively a node in the domain(1) and domain(n). Path Key confidentiality as per RFC5520 [RFC5520] SHOULD be used to obfuscate the detailed ERO(i) of the different domains(i). The resulting ERO is of the form (S, PKS(1), BN-ex(1), ..., BN-en(i), PKS(i), BN-ex(i),

..., BN-en(n), PKS(n), D) when Path Key is used and of the form {S, R(1,1), ..., R(1,k), BN-ex(1), ..., BN-en(i), R(i,1), ..., R(i,l), BN-ex(i), ..., BN-en(n), R(n,1), ..., R(n,m), D} otherwise.

The complete procedure with Path Key follow the different steps described below:

Step 1: Initialization

1. The Parent PCE sends a PCInitiate message to Child PCE(n) with an ERO = {PKS(n)} and End-Points = {BN-en(n), D}. Then, PCE(n) retrieves the ERO from the PKS(n) (if necessary) and sends to BN-en(n) a PCInitiate message with the ERO(n) = {BN-en(n), ..., D}, PST = TBD2 and End-Points Object = {BN-en(n), D} in order to inform the PCC BN-en(n) that this local path(n) is part of an inter-domain path.

2. When the PCC BN-en(n) receives the PCInitiate message from its PCE(n), it sets up the local path from the entry BN-en(n) to D by means of RSVP-TE signaling with the given ERO(n).

3. Once the path is set up, it chooses a free label for the Stitching Label SL(n) and adds a new entry in its MPLS L(F)IB with this SL(n) label. Then, it sends a PCRpt message to its PCE(n) with an RRO equal to {[LK(n), SL(n)], RRO(n)} and PLSP-ID(n).

4. Once PCE(n) receives the PCRpt from the PCC BN-en(n) with the RRO, PLSP-ID and PST = TBD2, it sends to its Parent PCE a PCRpt containing the RRO equal to {[LK(n), SL(n)]} and PLSP-ID(n). PCE(n) MAY add PKS(n) in the RRO.

Steps i: Actions performed for all intermediate domains(i), for i = n-1 to 2

1. The Parent PCE sends a PCInitiate message to Child PCE(i) with PST = TBD1, ERO = {PKS(i), [LK(i+1), SL(i+1)]} and End-Points = {BN-en(i), BN-ex(i)}

2. Then, PCE(i) retrieves the ERO from the PKS(i) if necessary and sends to the PCC BN-en(i) a PCInitiate message with ERO = {ERO(i), [LK(i+1), SL(i+1)]}, PST = TBD2 and End-Points Object = {BN-en(i), BN-ex(i)} in order to inform the PCC BN-en(i) that this local path(i) is part of an inter-domain path.

3. When the PCC BN-en(i) receives the PCInitiate message from its PCE(i), it sets up the local path from BN-en(i) to BN-ex(i) by means of RSVP-TE signaling with the given ERO(i).

4. Egress Control mechanism, as per RFC4003 section 2.1 [RFC4003], is used to instruct the egress node of domain(i), i.e. BN-ex(i) to forward packets belonging to this tunnel with the Stitching Label. Both the Label Stitching and an identifier of the outgoing interface are carried in the ERO = {..., [LK(i+1), SL(i+1)]} as the last SubObject in conformance to [RFC4003]. So that, BN-ex(i) installs in its MPLS L(F)IB the SWAP instruction to label SL(i+1) with forward to LK(i+1) instead of the usual POP instruction.

5. Once the tunnel is set up, PCC BN-en(i) chooses a free label for the Stitching Label SL(i) and adds a new entry in its MPLS L(F)IB with this SL(i) label. Then, it sends a PCRpt message to its PCE(i) with an RRO equal to {[LK(i), SL(i)], RRO(i)} and PLSP-ID(i).

6. Once PCE(i) receives the PCRpt from the PCC BN-en(i) with the RRO and PST = TBD2, it sends to its Parent PCE a PCRpt message containing the RRO equal to {[LK(i), SL(i)]} and the PLSP-ID(i). PCE(i) MAY add PKS(i) in the RRO.

7. Once the Parent PCE receives the PCRpt from the Child PCE(i), it stores the corresponding PLSP-ID for this inter-domain path part.

Steps n: Actions performed to the source domain(1)

Finally, the Parent PCE sends a last PCInitiate message to its Child PCE(1) with PST = TBD1, ERO = {PKS(1), [LK(2), SL(2)]} and End-Points = {S, BN-ex(1)}. In turn, Child PCE(1) sends a PCInitiate message to PCC node S with ERO equal to {ERO(1), [LK(2), SL(2)]}, PST = 0 and End-Points Object = {S, BN-ex(1)}. This time, the PST is equal to 0 as the PCC S does not need to return a Stitching Label SL, because it is the head-end of the inter-domain path. A usual PCRpt message is sent back to PCE(1) by the PCC node S. In turn, Child PCE(1) sends a final PCRpt message to the Parent PCE with the PSLP-ID(1). PCE(1) MAY add {S, BN-ex(1)} in the RRO as a loose path.

4.2. Completion Failure of Inter-domain Path Setup Procedure

In case of error during path set up, PCRpt and or PCError messages MUST be used to signal the problem to the Parent PCE. In particular, if the new PST values defined in this document are not supported by the Child PCE or the PCC, the Child PCE, respectively the PCC, MUST return a PCErr message with Error-Type = 21 (TE path setup error) and Error-Value = 1 (Unsupported path setup type) to its Parent PCE. If Child PCE(i) receives a PCInitiate message from its Parent PCE without PST set to TBD1 or PST set to a value different from TBD1, it MUST return a PCErr message with Error-Type = 21 (TE path setup

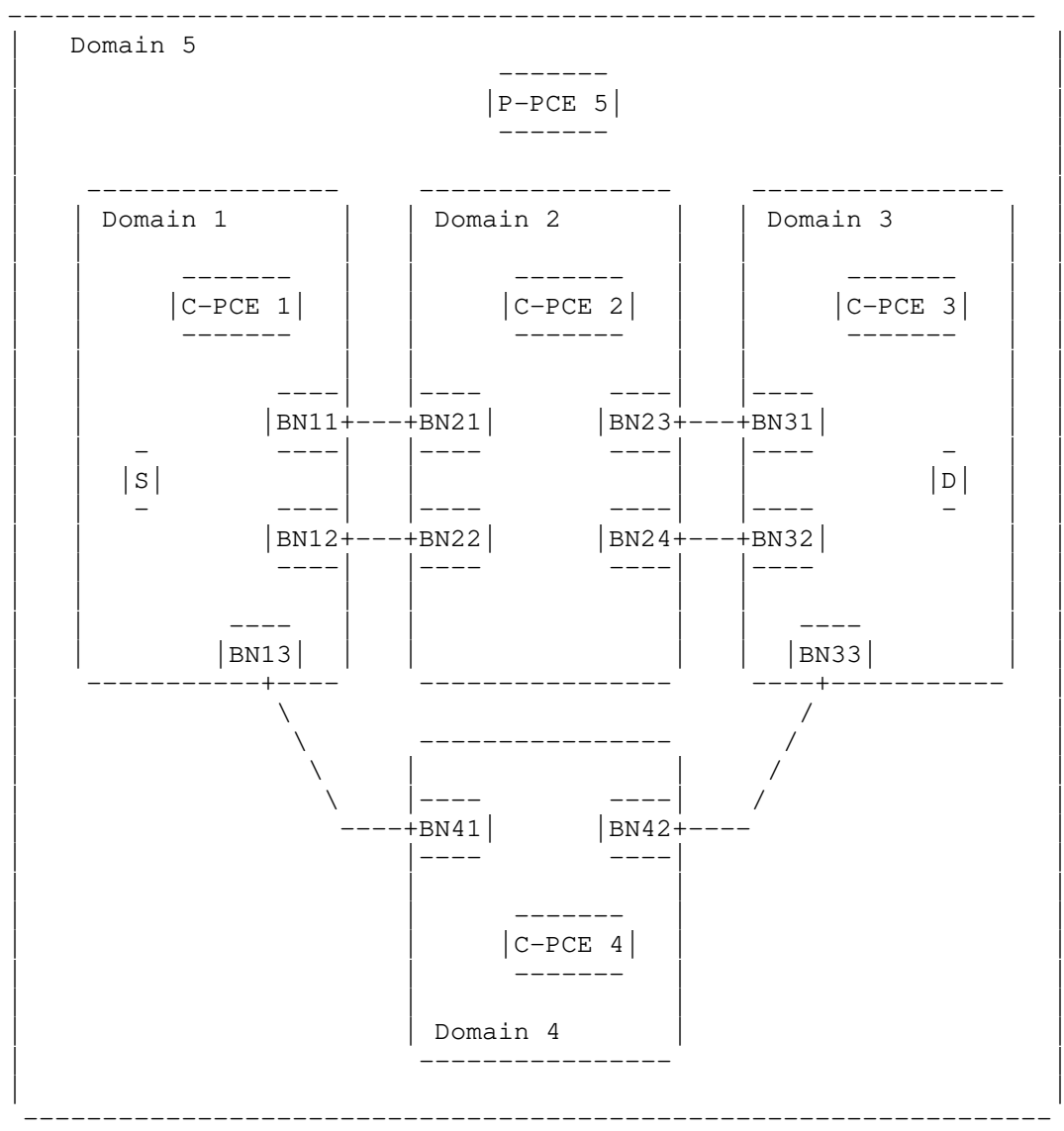
error) and Error-Value = 1 (Unsupported path setup type) to its Parent PCE.

Following a PCInitiate message with PST set to TBD1, if a Child PCE or a PCC returns no RRO, or an RRO without the Stitching Label SL and an identifier of the associated link, the Parent PCE, respectively the Child PCE, MUST return a PCErr message with Error-Type = 21 (TE path setup error) and Error-Value = TBD5 (Mandatory Stitching Label missing in the RRO).

In case of completion failure, the Parent PCE MUST send a PCInitate message (R flag set in the SRP Object as per [RFC8281]) to tear down this inter-domain path from the Child PCEs that already set up their respective part of the inter-domain path. Child PCE(i) MUST remove its local path by means of PCInitiate message with R flag set to 1 to its PCC BN-en(i) and send back a PCRpt message to the Parent PCE.

4.3. Example for Stateful H-PCE Stiching Procedure

Taking the sample hierarchical domain topology example from [RFC6805] as the reference topology for the entirety of this section.



Hierarchical domain topology from RFC6805

Section 3.3.1 of [RFC8751] describes the per-domain stitched LSP mode and list all the steps needed. To support SL-based stitching, using the reference architecture described in the figure above, the steps are modified as follows (note that we do not use PKS in this example for simplicity):

Step 1: initialization

The P-PCE (PCE5) is requested to initiate a path. Steps 4 to 10 of section 4.6.2 of [RFC6805] are executed to determine the end-to-end path, which are split into per-domain paths, e.g. {S-BN41, BN41-BN33, BN33-D}.

Step 2: Path (BN33-D) at C-PCE3:

1. The P-PCE (P-PCE5) sends the initiate request to the C-PCE (C-PCE3) via PCInitiate message for path (BN33-D) with ERO={BN33..D} and PST = TBD1.

2. C-PCE3 further propagates the initiate message to BN33 with the ERO and PST = TBD2/TBD3 based on the setup type.

3. BN33 initiates the setup of the path and reports to the status ("GOING-UP") to C-PCE3.

4. C-PCE3 further reports the status of the path to the P-PCE (P-PCE5)

5. The node BN33 notifies the path state to C-PCE3 when the state is "UP"; it also sends the Stitching Label (SL33) in the RRO as {SL33,BN33..D}.

6. C-PCE3 further reports the status of the path to the P-PCE (P-PCE5) as well as sends the Stitching Label (SL33) in the RRO as {LK33,SL33,BN33..D}.

Step 3: Path (BN41-BN33) at C-PCE4

1. The P-PCE (P-PCE5) sends the initiate request to the C-PCE (C-PCE4) via PCInitiate message for path (BN41-BN33) with ERO={BN41..BN42,LK33,SL33,BN33} and PST = TBD1.

2. C-PCE4 further propagates the initiate message to BN41 with the ERO and PST = TBD2/TBD3 based on the setup type. In case of RSVP_TE, the node BN41 encode the Stitching Label SL33 as part of the ERO to make sure the node BN42 uses the label SL33 towards node BN33. In case of SR, the label SL33 is part of the label stack pushed at node BN41.

3. BN41 initiates the setup of the path and reports the path status ("GOING-UP") to C-PCE4.

4. C-PCE4 further reports the status of the path to the P-PCE (P-PCE5).

5. The node BN41 notifies the path state to C-PCE4 when the state is "UP"; it also sends the Stitching Label (SL41) in RRO as {LK41,SL41,BN41..BN33}.

6. C-PCE4 further reports the status of the to the P-PCE (P-PCE5) as well as sends the Stitching Label (SL41) in the RRO as {LK41,SL41,BN41..BN33}.

Step 3: Path (S-BN41) at C-PCE1

1. The P-PCE (P-PCE5) sends the initiate request to the C-PCE (C-PCE1) via PCInitiate message for path (S-BN41) with ERO={S..BN13,LK41,SL41,BN41}.

2. C-PCE1 further propagates the initiate message to node S with the ERO. In case of RSVP-TE, node S encodes the Stitching Label SL41 as part of the ERO to make sure the node BN13 uses the label SL41 towards node BN41. In case of SR, the label SL41 is part of the label stack pushed at node S.

3. S initiates the setup of the path and reports the path status ("GOING-UP") to C-PCE1.

4. C-PCE1 further reports the status of the path to the P-PCE (P-PCE5)

5. The node S notifies the path state to C-PCE1 when the state is "UP".

6. C-PCE1 further reports the status of the path to the P-PCE (P-PCE5).

In this way, per-domain paths are stitched together using the Stitching Label (SL). The per-domain paths MUST be set up from the destination domain towards the source domain one after the other.

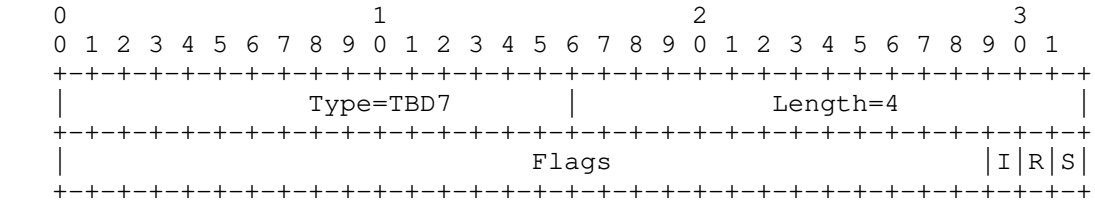
Once the per-domain path is set up, the entry BN chooses a free label for the Stitching Label SL and adds a new entry in its MPLS L(F)IB with this SL label. The SL from the destination domain is propagated to adjacent transit domain, towards the source domain at each step. This happens from the entry BN to C-PCE then to the P-PCE, and vice-versa. In case of RSVP-TE, the entry BN further propagates the SL label to the exit BN via RSVP-TE. In case of SR, the SL label is pushed as part of the SR label stack.

5. Inter-domain Path Management

This section describes how inter-domain paths could be managed.

5.1. Stitching Label PCE Capabilities

A PCE needs to know if its neighbor PCEs as well as PCCs are able to configure and provide a Stitching Label. The STITCHING-LABEL-PCE-CAPABILITY TLV is an optional TLV for use in the OPEN object for Stitching Label PCE capability advertisement. Its format is shown in the following figure:



STITCHING-LABEL-PCE-CAPABILITY TLV Format

The Type (16 bits) of the TLV is TBD7. The Length field is 16 bits long and has a fixed value of 4.

The value comprises a single 32 bits "Flags" field:

R (RSVP-TE-STITCHING-LABEL-CAPABILITY - 1 bit): if set to 1 by a PCC, the R flag indicates that the PCC is able to provide Stitching Labels, for RSVP-TE inter-domain paths, when requested by a PCE. If set to 1 by a PCE, the R flag indicates that the domain controlled by this PCE is able to set up inter-domain paths by means of RSVP-TE signaling.

S (SEGMENT-ROUTING-STITCHING-LABEL-CAPABILITY - 1 bit): if set to 1 by a PCC, the S flag indicates that the PCC is able to provide Stitching Labels, for Segment-Routing inter-domain paths, when requested by a PCE. If set to 1 by a PCE, the R flag indicates that the domain controlled by this PCE is able to set up inter-domain paths by means of Segment Routing.

I (INTER-DOMAIN-STITCHING-LABEL-CAPABILITY - 1 bit): if set to 1 by a PCE, the I flag indicates that the domain is supporting Stitching Label to set up inter-domain paths. This flag is reserved for PCEP session established between PCEs and MUST be kept unset by a PCC.

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

PCCs MUST set the R and/or S flags and MUST NOT set the I flag when adding the Stitching Label Capability to the PCEP Open Message. The RSVP-TE-STITCHING-LABEL-CAPABILITY, respectively SEGMENT-ROUTING-STITCHING-LABEL-CAPABILITY, flag must be set by both the PCC and PCE in order to enable the configuration of Stitching Labels with RSVP-TE, respectively with Segment-Routing.

A PCE MUST set the I flag when establishing a PCEP session with a neighbor PCE when adding Stitching Label Capability to the PCEP Open Message. It MAY set R and/or S flags depending if the operator would like to keep confidential the technology used to set up inter-domain paths or not. The INTER-DOMAIN-STITCHING-LABEL-CAPABILITY flag must be set by both PCEs in order to enable inter-domain paths instantiation by means of Stitching Label.

5.2. Identification of Inter-domain Paths

First, in order to manage inter-domain paths composed by the stitching or nesting of local paths, it is important to identify them. For this purpose, the PLSP-ID managed by the PCEs are combined to one provided by PCCs to form a global identifier as follow:

- PCE(i) in the Backward Recursive method or the Child PCE in Hierarchical method MUST create a new unique PLSP-ID for this inter-domain path part and MUST send it in the PCRpt message, to the PCE(i-1), respectively the Parent PCE. In addition this new PLSP-ID MUST be associated to the one received from the PCC that instantiates the local path part for further reference.

- In the Hierarchical mode, the Parent PCE MUST store and associate the different PLSP-ID(i)s received from the different Child PCE(i)s in order to identify the different part of the inter-domain paths.

- In the Backward Recursive method, PCE(i) MUST store and associate its PLSP-ID(i) and the PLSP-ID(i+1) it received from the PCE(i+1). PCE(n), i.e. the last one in the chain, does not need to perform such association.

Further reference to the inter-domain path will use this PLSP-ID(i). In the Backward Recursive method, PCE(i) MUST replace the PLSP-ID(i) by PLSP-ID(i+1) in the PCUpd, PCRpt or PCinitiate message before propagating it to PCE(i+1); and PCE(i) MUST replace the PLSP-ID(i+1) by PLSP-ID(i) in the PCRpt message before propagating it to the

PCE(i-1). In the Hierarchical method, the Parent PCE MUST use the corresponding PLSP-ID(i) of the Child PCE(i).

5.3. Inter-domain Association Group

In case of failure, a PCE(i) will received PCRpt messages from its PCCs and neighbors PCE(i+1) to synchronize the Inter-domain paths. In addition, it may received PCInitiate messages from its previous neighbors PCE(i-1) to re-initiate its inter-domain path part. As the PCE(i) may loose the PLSP-ID association, a new association group (within Association Object) is used to ease the association of the different parts of the inter-domain path: the local part and the PCE-to-PCE part. The use of the Association Object is MANDATORY in the Backward Recursive method and OPTIONAL in the Hierarchical method.

For that purpose, a new Inter-Domain Association Type with value TBD4 is defined. The first PCE in the Backward Recursive chain (the one which received the initial request) MUST send the PCInitiate message with an Association Object as follows:

- Association Type field MUST be set to new value TBD4
- Association ID MUST be set to a unique value. In case the Association ID field is too short or wraps, the first PCE MAY use the Extended Association ID to increase the number of association groups. The Association ID is managed locally by the PCE and does not need to be coordinated with neighbor or remote PCEs.
- IPV4 or IPv6 association source MUST be set to the IP address which identifies PCE(1) in domain(1).
- The Global Association Source TLV MUST be present and set with the ASN number of domain(1). It allows to create a globally unique association scope without putting constraint on operator's IP association source. Thus the IP Association Source is associated with the Global Association source to form a unique identifier.
- Extended Association ID MAY be present and MANDATORY if association ID is too short or wraps.

Subsequent PCE(i), for i = 2 to n, MUST send this Association Object as is to the local PCC and the neighbor PCE(i+1).

In case of error with the association group, a PCErr message MUST be raised with Error = 26 (Association Error) and Error value set accordingly. A new Error value TBD6 is defined to identify association of inter-domain paths.

In the Hierarchical method, the Parent PCE MAY act as the initiator of the Association and send to the Child PCEs an Association Object that follows the same rules as for the Backward Recursive method. In turn, Child PCEs MUST propagate the Association Object to the local PCCs as is.

5.4. Modification of Inter-domain Paths

For the Backward Recursive method, each domain manages their respective local path part of an inter-domain path independently of each other. In particular, Stitching Label(i) is managed by domain(i) and is of interest of domain(i-1) only. Thus, Stitching Label SL(i) is not supposed to be propagated to other domains. The same behavior apply to PLSP-ID(i). In the Hierarchical method, the Parent PCE MUST ensure the correct distribution of Stitching Label SL(i) to Child PCE(i-1). The PLSP-ID(i) is kept for the usage of the Parent PCE and thus is not propagated. Only the Association Object defined in section 5.2 is propagated if it is present.

If PCE(i) needs to modify its local path(i) with a PCUpd message to the PCC BN-en(i), once the PCRpt message received from the PCC BN-en(i), it MUST sends a new PCRpt message to advertise the modification. This message is targeted to its neighbor PCE(i-1) in the Backward Recursive method, respectively to the Parent PCE in the Hierarchical method. In this case PLSP-ID(i) is used to identify the inter-domain path. PCE(i-1), respectively the Parent PCE, MUST propagate the PCRpt message if the modification implies the upstream domain, e.g. if the PCRpt indicates that the Stitching Label SL(i) has changed.

PCE(1), respectively the Parent PCE, could modify the inter-domain path. For that purpose, it MUST send a PCUpd message to its neighbor PCEs, respectively Child PCE, using the PLSP-ID it received. Each PCE(i) MUST process the PCUpd message the same way they process the PCInitiate message as define in section 3.1 for the Backward Recursive method and in section 4.1 for the Hierarchical method.

In case a failure appear in domain(i), e.g. path becoming down, PCE(i) MUST sends a PCRpt message to its neighbor PCE(i-1), respectively its Parent PCE to advertise the problem in its local part of the inter-domain path. Once PCE(1), respectively the Parent PCE, receives this PCRpt message indicating that the path is down, it is up to the PCE(1), respectively the Parent PCE to take appropriate correction e.g. start a new path computation to update the ERO.

5.5. Modification of Inter-domain Paths

Modification of local path, BN-en(i) and BN-ex(i) is left for further study.

5.6. Tear-Down of Inter-domain Paths

The tear-down of an inter-domain path is only possible by the inter-domain path initiator i.e. PCE(1). For the Backward Recursive method, a PCInitiate message with R flag set to 1, PLSP-ID set accordingly to section 5.1 and the Association Object with R flag set to 1, is sent by PCE(1) to PCE(n) through PCE(i), and processed the same way as described in section 3.1. For the Hierarchical method, a PCInitiate message with R flag set to 1 is sent by the Parent PCE to each Child PCE(i) with corresponding PLSP-ID(i), and processed according to section 4.1. Each domain PCE(i) is responsible to tear down its part of the path and the PCC MUST release both the Stitching label SL in its L(F)IB and the path when it receives the PCInitiate message with the R flag set to 1 and the corresponding PLSP-ID. The Association Group MUST also be removed by the PCC and PCE(i).

6. Applicability

The newly introduce Stitching Label SL serves to stitch or nest part of local paths to form an inter-domain path. Each domain is free to decide if the incoming path is stitched or nested and how the path is enforced, e.g. through RSVP-TE or Segment Routing. At the peering point, the Border Node BN-ex(i) MUST encapsulate the packet with the Stitching Label, i.e. the MPLS label prior to send them to the next Border Node BN-en(i+1). Thus, only RSVP-TE and Segment Routing over MPLS technology are detailed in the following sections.

6.1. RSVP-TE

In case of RSVP-TE, the Border Node BN-ex(i) needs to received the Stitching Label from BN-en(i) through the RSVP-TE message and install in its L(F)IB a SWAP instruction to the Stitching Label and forward it to the next Border Node BN-en(i+1). For that purpose, the Egress Control mechanism, as per RFC4003 section 2.1 [RFC4003], is RECOMMENDED to instruct the Border Node BN-ex(i) of this action. Other mechanisms to program the L(F)IB could be used, e.g. NETCONF.

As the Stitching Label could serves to stitch or nest tunnels, a domain(i) may decide to nest the incoming LSPs into a higher hierarchy of LSPs for a Traffic Engineering purpose. A PCE(i) may also decide to group local LSPs part of inter-domain paths into a higher hierarchical LSP to carry all these local paths from a BN-en(i) to a BN-ex(i).

6.2. Segment Routing

To use Segment Routing instead of RSVP-TE to set up the local LSP tunnels as defined in [RFC8664], PCE(i) MUST send a PCInitiate message with PST = TBD3 instead of TBD2 to advertise its respective PCC that the local path is enforce by means of Segment Routing.

The Stitching Label SL(i+1) will be inserted into the label stack in order to become the top label in the stack when the packet reaches BN-en(i+1). Thus, the Stitching Label SL(i+1) serves as a FEC entry for BN-en(i+1) to identify the packets that follow the next Segment Path. For that purpose, BN-en(i+1) MUST install in its MPLS L(F)IB an instruction to replace the incoming Stitching Label SL(i+1) by the label stack given by the ERO(i+1) plus the Stitching Label SL(i+2), if any.

When a packet reaches BN-ex(i), the last label in the stack before the label SL(i+1) corresponds to a SID that allows to reach BN-en(i+1). When there are multiple interfaces between Border Nodes, BN-ex(i) needs to know how to send the packets to BN-en(i+1). Similarly to the Egress Control mechanism used with RSVP-TE, it is RECOMMENDED to use the inter-domain SID defined as per draft Egress Peer Engineering [I-D.ietf-idr-bgpls-segment-routing-epe] for that purpose. The inter-domain SID is announced by BN-ex(i) to PCE(i) through BGP-LS for each interface that connect BN-ex(i) to neighbors BN-en(i+1). Thus, the label stack will end with {BN-ex(i) SID, Inter-Domain SID, SL(i+1)} and should be processed as follows:

- The penultimate router of domain(i) pops its node SID, and sends the packet to the next node designated by the top label in the label stack, i.e. the node SID of BN-ex(i) or the adjacency SID of the link between the router and BN-ex(i).

- BN-ex(i) pops its node SID or its adjacency SID and looks up the next label in the stack, i.e. the inter-domain SID which corresponds to the interface to BN-en(i+1). BN-ex(i) pops this inter-domain SID as well and sends the packet to BN-ex(i) through the corresponding interface.

- BN-en(i+1) looks up the top label which is the Stitching Label SL(i+1), pops it and replaces it by the sub-sequent label stack.

Other mechanisms, e.g. NETCONF, could be used to configure the inter-domain SID on exit Border Nodes.

6.3. Mixing Technologies

During the instantiation procedure, if PCE(i) decides to reuse a local tunnel which is not yet part of an inter-domain tunnel, it SHOULD send a PCUpd message with PST = TBD2 to the PCC BN-en(i), in order to request a Stitching Label SL(i), and new ERO(i) to add the Stitching Label SL(i+1) and the associated link to the previous ERO.

[RFC8453] describes framework for Abstraction and Control of TE Networks (ACTN), where each Physical Network Controller (PNC) is equivalent to C-PCE and the Multi-Domain Service Coordinator (MDSC) to the P-PCE. The per-domain stitched LSP as per the Hierarchical PCE architecture described in Section 3.3.1 and Section 4.1 of [RFC8751] is well suited for ACTN. The Stitching Label mechanism as described in this document is well suited for ACTN when per-domain LSPs need to be stitched to form an E2E tunnel or a VN Member. It is to be noted that certain VNs require isolation from other clients. The SL mechanism described in this document can be applicable to the VN isolation use-case by uniquely identifying the concatenated stitching labels across multi-domain only to a certain VN member or an E2E tunnel.

As each operator is free to enforce the tunnel with its technology choice, it is a local policy decision for PCE(i) to instantiate the local part of the end-to-end tunnel by either RSVP-TE or Segment Routing. Thus, the PST value (i.e. TBD2 or TBD3) used in the PCinitiate message sent by the PCE(i) to the local PCC is determined by the local policy. How the local policy decision is set in the PCE is out of the scope of this document. This flexibility is allowed because the SL principle allows to mix (data plane) technologies between domains. For example, a domain(i) could use RSVP-TE while domain(i+1) uses SR. The SL could serve to stitch indifferently Segment Paths and RSVP-TE tunnels. Indeed, the SL will be part of the label stack in order to become the top label in the stack when reaching the BN-en(i+1). This SL could be swapped as usual if the next domain uses RSVP-TE tunnels. When the upstream domain uses an RSVP-TE tunnel, the SL will serve as a key for the BN-en(i+1) to determine which label stack it must use on top of the packet for a Segment Routing path.

6.4. Inter-Area

If use cases for inter-AS are easily identifiable, this is less evident for inter-area. However, two scenarios have been identified:

- Paths between levels for IS-IS networks.
- Reduction of labels stack depth for Segment Routing.

Thus, the SL could be used to stitch or nest independent tunnels deployed through different IS-IS levels, even if there are controlled by the same PCE. IS-IS levels are considered as domains but under the control of the same PCE. In this scenario, there is no exchange between PCEs (it remains internal and implementation matter) and new TLVs are only applicable between the PCE and PCCs. The PCE requests to the different PCCs it identifies (i.e. BNs of the different IS-IS levels) to set up SLs and propagated them.

In large-scale networks, MSD could constraints the path computation in the possibility of path selection i.e. explicit expression of a path could exceeded the MSD. The SL could be used to split a too long explicit path regarding the MSD constraints. In this scenario, there is also no communications between PCEs and new TLVs are only used between PCE and PCCs.

7. IANA Considerations

7.1. Path Setup Type Values

[RFC8408] defines the PATH-SETUP-TYPE TLV. IANA is requested to allocate new code points in the PCEP PATH-SETUP-TYPE TLV PST field registry, as follows:

Value	Description	Reference
TBD1	Inter-domain TE end-to-end path is set up using the Backward Recursive method	This Document
TBD2	Inter-domain TE local path is set up using RSVP-TE signaling	This Document
TBD3	Inter-domain TE local path is set up using Segment Routing	This Document

7.2. Association Type Value

PCE Association Group [RFC8697] defines the ASSOCIATION Object and requests that IANA creates a registry to manage the value of the Association Type value. IANA is requested to allocate a new code point in the PCEP ASSOCIATION GROUP TLV Association Type field registry, as follows:

Association Type	Description
TBD4	Inter-domain Association Group

7.3. PCEP Error Values

IANA is requested to allocate code-points in the PCEP-ERROR Object Error Values registry for a new error-value of Error-Type 21 Invalid TE path setup and new error-value of Error-Type 26 Association Error:

Error-Type	Error-Value	Description
21	TBD5	Mandatory Stitching Label missing in the RRO
26	TBD6	Error in association of Inter-domain LSPs

7.4. PCEP TLV Type Indicators

IANA is requested to allocate a new TLV Type Indicator for the "Stitching Label PCE Capability" within the "PCEP TLV Type Indicators" subregistry of the "Path Computation Element Protocol (PCEP) Numbers" registry:

Value	Description	Reference
TBD7	STITCHING-LABEL-PCE-CAPABILITY	This Document

7.5. Stitching Label PCE Capability

IANA is requested to allocate a new subregistry, named "STITCHING-LABEL-PCE-CAPABILITY TLV Flag Field", within the "Path Computation Element Protocol (PCEP) Numbers" registry, to manage the Flag field in the STITCHING-LABEL-PCE-CAPABILITY TLV of the PCEP OPEN object (class = 1). New values are 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

Value	Description	Reference
31	RSVP-TE-STITCHING-CAPABILITY	This Document
30	SEGMENT-ROUTING-STITCHING-CAPABILITY	This Document
29	INTER-DOMAIN-STITCHING-CAPABILITY	This Document

8. Security Considerations

No modification of PCE protocol (PCEP) has been requested by this draft which does not introduce any issue regarding security. Concerning the PCEP session between PCEs, authors recommend to use the secured version of PCEP as defined in PCEPS [RFC8253] or use any other secured tunnel mechanism, e.g. IPsec tunnel to transport PCEP session between PCEs.

9. Acknowledgements

The authors want to thanks PCE's WG members, and in particular Dhruv Dhody who greatly contributed to the Hierarchical section of this document and Quan Xiong for his advice.

10. Disclaimer

This work has been performed in the framework of the H2020-ICT-2014 project 5GEx (Grant Agreement no. 671636), which is partially funded by the European Commission. This information reflects the consortium's view, but neither the consortium nor the European Commission are liable for any use that may be done of the information contained therein.

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>.

[RFC5441] Vasseur, JP., Ed., Zhang, R., Bitar, N., and JL. Le Roux, "A Backward-Recursive PCE-Based Computation (BRPC) Procedure to Compute Shortest Constrained Inter-Domain Traffic Engineering Label Switched Paths", RFC 5441, DOI 10.17487/RFC5441, April 2009, <https://www.rfc-editor.org/info/rfc5441>.

[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>.

[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>.

11.2. Informative References

[I-D.ietf-idr-bgpls-segment-routing-epe] Previdi, S., Talaulikar, K., Filsfils, C., Patel, K., Ray, S., and J. Dong, "BGP-LS extensions for Segment Routing BGP Egress Peer Engineering", draft-ietf-idr-bgpls-segment-routing-epe-19 (work in progress), May 2019.

[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>.

[RFC3473] Berger, L., Ed., "Generalized Multi-Protocol Label Switching (GMPLS) Signaling Resource ReserVation Protocol-Traffic Engineering (RSVP-TE) Extensions", RFC 3473, DOI 10.17487/RFC3473, January 2003, <https://www.rfc-editor.org/info/rfc3473>.

[RFC4003] Berger, L., "GMPLS Signaling Procedure for Egress Control", RFC 4003, DOI 10.17487/RFC4003, February 2005, <https://www.rfc-editor.org/info/rfc4003>.

[RFC4206] Kompella, K. and Y. Rekhter, "Label Switched Paths (LSP) Hierarchy with Generalized Multi-Protocol Label Switching (GMPLS) Traffic Engineering (TE)", RFC 4206, DOI 10.17487/RFC4206, October 2005, <https://www.rfc-editor.org/info/rfc4206>.

[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>.

[RFC5150] Ayyangar, A., Kompella, K., Vasseur, JP., and A. Farrel, "Label Switched Path Stitching with Generalized Multiprotocol Label Switching Traffic Engineering (GMPLS TE)", RFC 5150, DOI 10.17487/RFC5150, February 2008, <https://www.rfc-editor.org/info/rfc5150>.

[RFC5520] Bradford, R., Ed., Vasseur, JP., and A. Farrel, "Preserving Topology Confidentiality in Inter-Domain Path Computation Using a Path-Key-Based Mechanism", RFC 5520, DOI 10.17487/RFC5520, April 2009, <https://www.rfc-editor.org/info/rfc5520>.

[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>.

[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>.

[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>.

[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>.

[RFC8751] Dhody, D., Lee, Y., Ceccarelli, D., Shin, J., and D. King, "Hierarchical Stateful Path Computation Element (PCE)", RFC 8751, DOI 10.17487/RFC8751, March 2020, <https://www.rfc-editor.org/info/rfc8751>.

Authors' Addresses

Olivier Dugeon
Orange Labs
2, Avenue Pierre Marzin
Lannion 22307
France

Email: olivier.dugeon@orange.com

Julien Meuric
Orange Labs
2, Avenue Pierre Marzin
Lannion 22307
France

Email: julien.meuric@orange.com

Young Lee
Samsung Electronics

Email: younglee.tx@gmail.com

Daniele Ceccarelli
Ericsson
Torshamnsgatan, 48
Stockholm
Sweden

Email: daniele.ceccarelli@ericsson.com

PCE Working Group
Internet-Draft
Intended status: Standards Track
Expires: August 20, 2021

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
February 16, 2021

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

Abstract

Current PCEP standards allow only one intended and/or actual path to be present in a PCEP report or update. Applications that require multipath support such as SR Policy require an extension to allow signaling multiple intended and/or actual paths within a single PCEP message. This document introduces such an extension. Encoding of multiple intended and/or actual paths is done by encoding multiple Explicit Route Objects (EROs) and/or multiple Record Route Objects (RROs). A special separator object is defined in this document, to facilitate this. This mechanism is applicable to SR-TE and RSVP-TE and is dataplane agnostic.

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 20, 2021.

Table of Contents

1. Introduction

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

PCEP 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 (containing multiple SR-ERO subobject), 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).

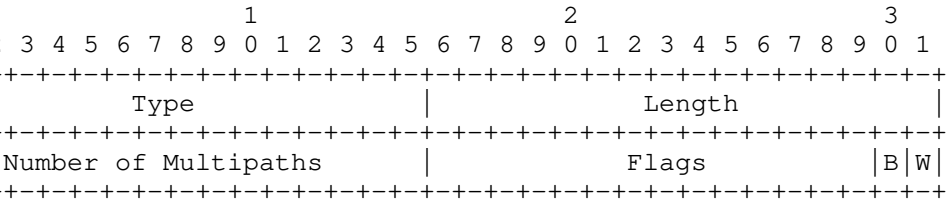


Figure 1: MULTIPATH-CAP TLV format

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.

Flags: Following bits are defined:

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

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

Unassigned bits are for future use. They MUST be set to 0 on transmission and MUST be ignored on receipt.

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.

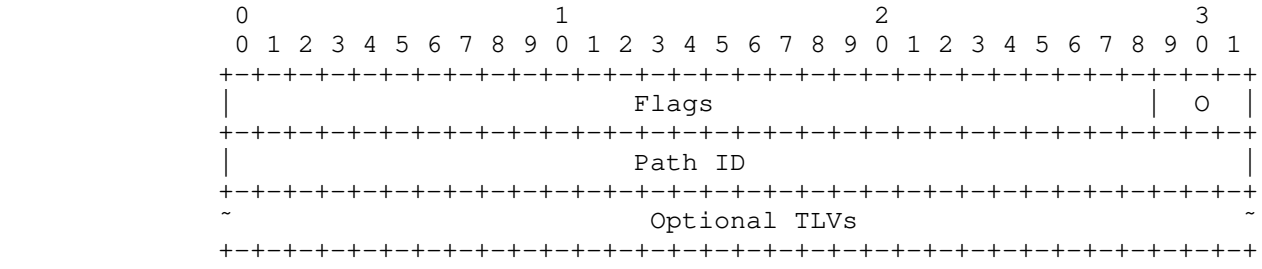


Figure 2: PATH-ATTRIB object format

Flags (32-bits): Following bits are assigned -

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

Unassigned bits are for future use. They MUST be set to 0 on transmission and MUST be ignored on receipt.

Path ID: 4-octet identifier that identifies a path in the set of multiple paths. It uniquely identifies a path (encoded in the ERO/RRO) within the set of multiple paths under the PCEP LSP. Once a path changes, a new Path ID is assigned.

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.

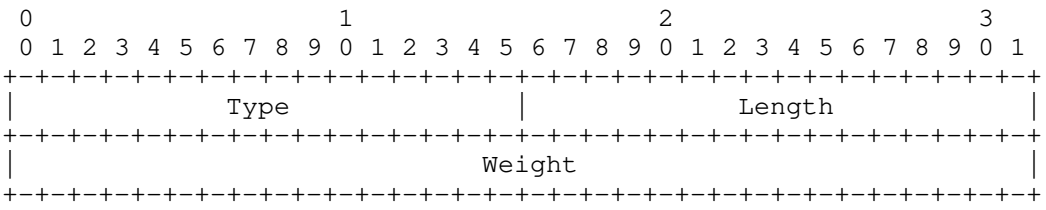


Figure 3: MULTIPATH-WEIGHT TLV format

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:

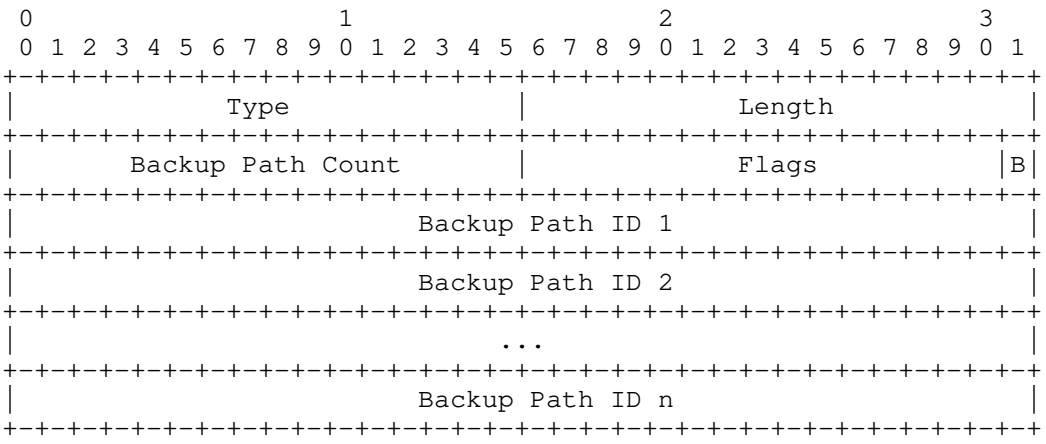


Figure 4: MULTIPATH-BACKUP TLV format

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).

Flags (16 bits): a flag field. Currently a single flag "B bit" is defined.
Unused flags MUST be set to zero while sending and ignored on receipt.

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. 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.peng-pce-te-constraints]. 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 SR-ERO/SR-RRO object MUST be included as per the existing RBNF, i.e., SR-ERO/SR-RRO MUST contain no sub-objects. If the head-end receives a non-empty SR-ERO/SR-RRO, 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

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 PCError 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>>
<SR-ERO (empty)>
<PATH-ATTRIB Path_ID=2 <WEIGHT-TLV Weight=W2> <COLOR-TLV Color=2>>
<SR-ERO (empty)>
```

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

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".

Following bits are defined:

Bit	Description	Reference
0-13	Unassigned	This document
14	B-flag: Backup support	This document
15	W-flag: Weighted ECMP support	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".

Following bits are defined:

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".

Following bits are defined:

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

12. References

12.1. Normative References

[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-02 (work in progress), January 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-09 (work in progress), November 2020.

[I-D.koldychev-pce-operational]
Koldychev, M., Sivabalan, S., Negi, M., Achaval, D., and H. Kotni, "PCEP Operational Clarification", draft-koldychev-pce-operational-02 (work in progress), August 2020.

[I-D.peng-pce-te-constraints]
Peng, S., Xiong, Q., and F. Qin, "PCE TE Constraints for Network Slicing", draft-peng-pce-te-constraints-04 (work in progress), August 2020.

[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., 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: August 23, 2021

M. Koldychev
Cisco Systems, Inc.
S. Sivabalan
Ciena Corporation
S. Peng
Huawei Technologies
D. Achaval
Nokia
H. Kotni
Juniper Networks, Inc

February 19, 2021

PCEP Operational Clarification
draft-koldychev-pce-operational-03

Abstract

This document is meant to provide better clarity about how PCEP operates and hence to facilitate better interoperability between different equipment vendors. The content of this document has been compiled based on the feedback from several multi-vendor interop exercises. Several constructs are introduced to facilitate this, such as the LSP-DB and the ASSO-DB.

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 August 23, 2021.

Table of Contents

1. Introduction

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

PCEP 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 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.

The PCEP protocol has evolved from a simple stateless model into a stateful model with more features being added. Due to subtle differences in interpretation of existing PCEP standards, it was found that networking equipment vendors often had to adjust their implementations, in order to interoperate. This informational document is meant to clarify these subtle differences and agree on a final model that all major vendors have agreed on and that all other vendors can adopt. This document applies to RSVP-TE and Segment-Routing.

2. Terminology

The following terminologies are used in this document:

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.

MBB: Make-Before-Break. A procedure during which the head-end of a traffic-engineered path wishes to move traffic to a new path without losing any traffic, by first "making" a new path and then "breaking" the old path.

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 optional TLVs based on the association types, that provides 'information' related to the association type.

ERO: Explicit Route Object is the path of the LSP encoded into a PCEP object. To represent an empty ERO object, i.e., without any subobjects, we use the notation "ERO={}". To represent an ERO object containing some given sequence of subobjects, we use the notation "ERO={A}".

3. PCEP LSP Database

We introduce the concept of the LSP-DB, as a database of actual LSP state in the network. This concept is not explicitly defined in [RFC8231], but is fully compatible with it. We use the LSP-DB to describe how certain actions are performed, because it is easier to define actions as a function of database state, rather than as a function of previously received messages. The structure and format of the LSP-DB MUST be common among all dataplane types (i.e., RSVP-TE/SR-TE/SRv6), all instantiation methods (i.e., PCC-initiated/PCE-initiated), all destination types (i.e., point-to-point/point-to-multipoint).

Note that we use the term "Tunnel" somewhat loosely here, to mean "the object identified by the PLSP-ID". It may or may not be an actual tunnel in the implementation. For example, working and protect paths can be implemented as one tunnel interface, but in PCEP we would refer to them as two different Tunnels, because they would have different PLSP-IDs.

Note that the term "LSP", which stands for "Label Switched Path", if taken too literally would restrict our discussion to MPLS dataplane only. In this document, we allow the term "LSP" to refer to any path, regardless of the dataplane format. So that an LSP can refer to MPLS and SRv6 dataplane paths.

3.1. Structure

[RFC8231] states that the LSP-IDENTIFIERS TLV contains the key that MUST be used to differentiate different LSPs during make before break procedure. We further clarify here that PCEP LSPs exist in a 2-tier structure. The top tier is the "Tunnel", identified by the PLSP-ID and/or SYMBOLIC-NAME, while the lower tier is the "LSP", identified

by the values in LSP-IDENTIFIERS TLV. A single Tunnel may contain multiple LSPs at the same time, i.e., a Tunnel is a container for LSPs. A Tunnel MUST have at least one LSP and when the last LSP is removed from the Tunnel, the Tunnel itself is removed.

3.2. Synchronization

The stateful PCE MUST maintain the PCE LSP-DB, which stores Tunnels and LSPs. The PCE LSP DB is only modified by PCRpt messages. No other PCEP message may modify the PCE LSP DB. The PCC MUST also maintain the PCC LSP DB, which it MUST synchronize with the PCE LSP DB by sending PCRpt messages.

The PCC adds/removes entries to/from its LSP-DB based on what LSPs it creates/destroys in the network. There can be many trigger types for updating the PCC LSP-DB, some examples include PCUpd messages, local computation on the PCC, local configuration on the PCC, etc. The trigger type does not affect the content of the PCC LSP-DB, i.e., the content of the PCC LSP-DB is updated identically regardless of the trigger type.

Whenever a PCC modifies an entry it its PCC LSP-DB, it MUST send a PCRpt message to the PCE (or multiple PCEs), to synchronize this change. Ensuring this synchronization is always in place allows one to define behavior as a function of LSP-DB state, instead of defining behavior as a function of what PCEP messages were sent or received.

The PCE MUST always act on the latest state of the PCE LSP DB. Note that this does not mean that the PCE cannot use information from outside of LSP-DB. For example, the PCE can use other mechanisms to collect traffic statistics and use them in the computation. However, these traffic statistics are not part of the LSP-DB, but only reference it.

The LSP-DB on both the PCC and the PCE only stores the actual state in the network, it does not store the desired state. For example, consider the case of PCE Initiated LSP, configured on the PCE. When the operator modifies the configuration of this LSP, that is a change in desired state. The actual state has not yet changed, so LSP-DB is not modified yet. The LSP-DB is only modified after the PCE sends PCInit/PCUpd message to the PCC and the PCC decides to act on that message. When the PCC acts on message, it would update its own PCC LSP DB and immediately send PCRpt to the PCE to synchronize the change. When the PCE receives the PCRpt msg, it updates its own PCE LSP DB. After this, the PCC LSP DB and PCE LSP DB are in sync.

3.3. Stateful Bringup

[RFC8231] in section 5.8.2, allows delegation of an LSP in operationally down state, but at the same time mandates the use of PCReq, before sending PCRpt. In this document, we would like to make it clear that sending PCReq is optional.

We shall refer to the process of sending PCReq before PCRpt as "stateless bringup". In reality, stateless bringup introduces overhead and is not possible to enforce from the PCE, because the stateless PCE is not supposed to keep any per-LSP state about previous PCReq messages. It was found that many vendors choose to ignore this requirement and send the PCRpt directly, without going through PCReq. This section will serve to explain and to validate this behavior.

Even though all the major vendors today are moving to the stateful PCE model, it does not deprecate the need for stateless PCEP. The key property of stateless PCEP is that PCReq messages MUST NOT modify the state of the PCE LSP-DB in any way. Therefore, PCReq messages are useful for many OAM ping/traceroute applications where the PCC wishes to probe the network without having any effect on the existing LSPs.

The PCC MAY delegate an empty LSP to the PCE and then wait for the PCE to send PCUpd, without sending PCReq. We shall refer to this process as "stateful bringup". The PCE MUST support the original stateless bringup, for backward compatibility purposes. Supporting stateful bringup should not require introducing any new behavior on the PCE, because as mentioned earlier, the PCE MUST NOT modify LSP-DB state based on PCReq messages. So whether the PCE has received a PCReq or not, it MUST process the PCRpt all the same.

An example of stateful bringup follows. In our example the PCC starts off by using LSP-ID of 0. The value 0 does not hold any special meaning, any other 16-bit value could have been used.

PCC has no LSP yet, but wants to establish a path. PCC sends PCRpt(R-FLAG=0, D-flag=1, OPER-FLAG=DOWN, PLSP-ID=100, LSP-ID=0, ERO={}).

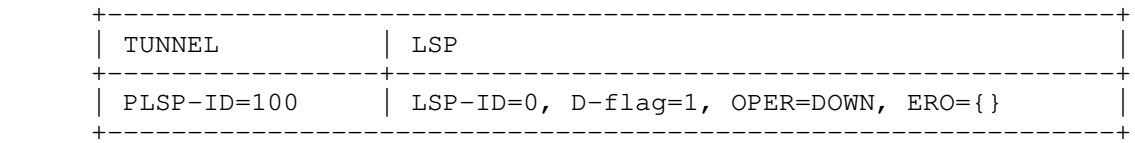


Figure 1: Content of LSP DB

PCC received a PCUpd from the PCE and has decided to install the ERO={A} from that PCUpd. PCC sends PCRpt(R-FLAG=0, D-flag=1, OPER-FLAG=UP, PLSP-ID=100, LSP-ID=0, ERO={A}).

TUNNEL	LSP
PLSP-ID=100	LSP-ID=0, D-flag=1, OPER=UP, ERO={A}

Figure 2: Content of LSP DB

3.4. Successful MBB

Below we give an example of doing MBB to switch the tunnel from one path to another. We represent the path encoded into the ERO object as ERO={A} and ERO={B}.

PCC has an existing LSP in UP state, with LSP-ID=2. PCC sends PCRpt(R-FLAG=0, PLSP-ID=100, LSP-ID=2, ERO={A}, OPER-FLAG=UP).

TUNNEL	LSP
PLSP-ID=100	LSP-ID=2, ERO={A}, OPER=UP

Figure 3: Content of LSP DB

PCC initiates the MBB procedure by creating a new LSP with LSP-ID=3. It does not matter what triggered the creation of the new LSP, it could have been due to a new path received via PCUpd (if the given tunnel is delegated), or it could have been local computation on the PCC (if the tunnel is locally computed on the PCC), or it could have been a change in configuration on the PCC (if the tunnel's path is explicitly configured on the PCC). It is important to emphasize that the procedure for updating the LSP-DB is common, regardless of the trigger that caused the change.

PCC sends PCRpt(R-FLAG=0, PLSP-ID=100, LSP-ID=3, ERO={B}, OPER-FLAG=UP).

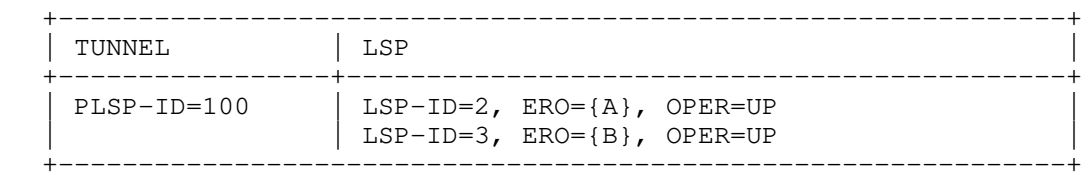


Figure 4: Content of LSP DB

After some time, the PCC decides to destroy the old LSP. PCC sends PCRpt(R-FLAG=1, PLSP-ID=100, LSP-ID=2).

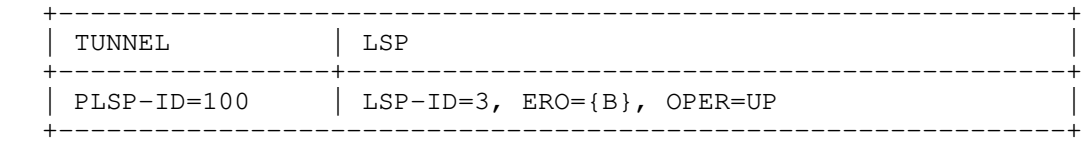


Figure 5: Content of LSP DB

3.5. Aborted MBB

The MBB process can abort when the newly created LSP is destroyed before it is installed as traffic carrying. This scenario is described below.

PCC has an existing LSP in UP state, with LSP-ID=2. PCC sends PCRpt(R-FLAG=0, OPER-FLAG=UP, PLSP-ID=100, LSP-ID=2).

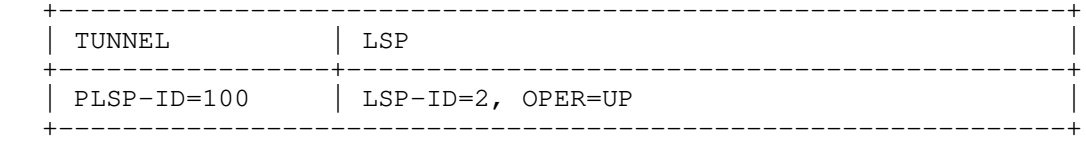


Figure 6: Content of LSP DB

MBB procedure is initiated, a new LSP is created with LSP-ID=3. LSP is currently being established, so its oper state is DOWN. PCC sends PCRpt(R-FLAG=0, OPER-FLAG=DOWN, PLSP-ID=100, LSP-ID=3).

TUNNEL	LSP
PLSP-ID=100	LSP-ID=2, OPER=UP LSP-ID=3, OPER=DOWN

Figure 7: Content of LSP DB

MBB procedure is aborted. PCC sends PCRpt(R-FLAG=1, PLSP-ID=100, LSP-ID=3).

TUNNEL	LSP
PLSP-ID=100	LSP-ID=2, OPER=UP

Figure 8: Content of LSP DB

4. PCEP Association Database

PCEP Association is a group of zero or more LSPs.

The PCE ASSO DB is populated by PCRpt messages and MAY also be populated via configuration on the PCE itself. An Association is identified by the Association Parameters. The Association parameters contain many fields, so for convenience we will group all the fields into a single value. We will use ASSO_PARAM=A, ASSO_PARAM=B, to refer to different PCEP Associations: A and B, respectively.

4.1. 2 LSPs in same Association

Below, we give an example of LSPs joining the same Association.

PCC creates the first LSP. PCC sends PCRpt(R-FLAG=0, PLSP-ID=100, LSP-ID=1, ASSO_PARAM=A, ASSO_R_FLAG=0).

ASSO	LSP
ASSO_PARAM=A	PLSP-ID=100, LSP-ID=1

Figure 9: Content of PCE ASSO DB

PCC creates the second LSP. PCC sends PCRpt(R-FLAG=0, PLSP-ID=200, LSP-ID=1, ASSO_PARAM=A, ASSO_R_FLAG=0).

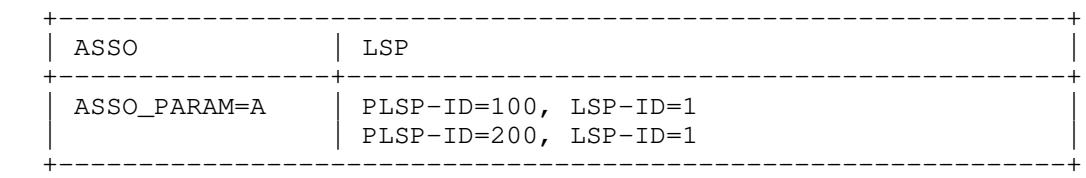


Figure 10: Content of PCE ASSO DB

PCC updates the first LSP, the PCC is NOT REQUIRED to send the ASSOCIATION object in this PCRpt, since the LSP is already in the Association. PCC sends PCRpt(R-FLAG=0, PLSP-ID=100, LSP-ID=1). The content of the PCE ASSO DB is unchanged. Note that the PCC MUST send the ASSOCIATION OBJECT in the first PCRpt during SYNC state, even if it has already issued a PCRpt with the association object sometime in the past with this PCE. The synchronization steps outlined in [RFC8697] are to be followed.

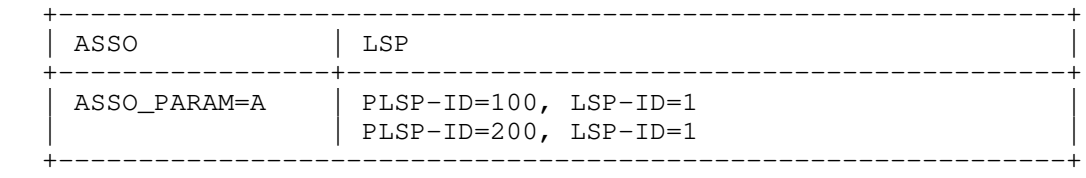


Figure 11: Content of PCE ASSO DB

PCC decides to delete the second LSP. PCC sends PCRpt(R-FLAG=1, PLSP-ID=200, LSP-ID=1).

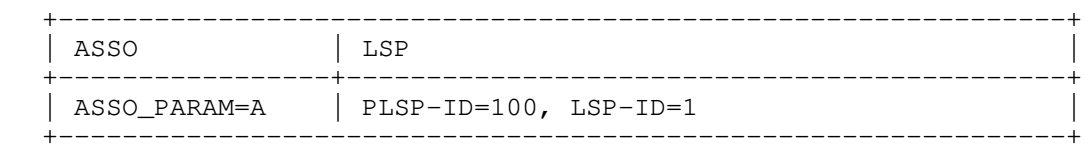


Figure 12: Content of PCE ASSO DB

PCC decides to remove the first LSP from the Association, but not delete the LSP itself. PCC sends PCRpt(R-FLAG=0, PLSP-ID=100, LSP-ID=1, ASSO_PARAM=A, ASSO_R_FLAG=1). The PCE ASSO DB is now empty.

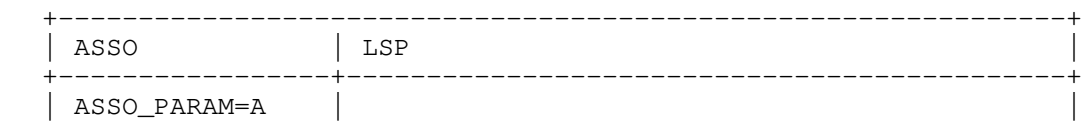


Figure 13: Content of PCE ASSO DB

4.2. Switch Association during MBB

Each new LSP (identified by the LSP-ID) does not inherit the Association membership of any previous LSPs within the same Tunnel. This is done so that a Tunnel can have two LSPs that are in different Associations, this may be required when switching from one Association to another.

Below, we give an example a Tunnel going through MBB and switching from Association A to Association B.

PCC creates the first LSP. PCC sends PCRpt(R-FLAG=0, PLSP-ID=100, LSP-ID=1, ASSO_PARAM=A, ASSO_R_FLAG=0).

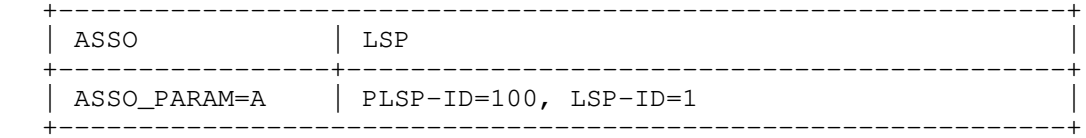


Figure 14: Content of PCE ASSO DB

PCC creates the MBB LSP in a different Association. PCC sends PCRpt(R-FLAG=0, PLSP-ID=100, LSP-ID=2, ASSO_PARAM=B, ASSO_R_FLAG=0).

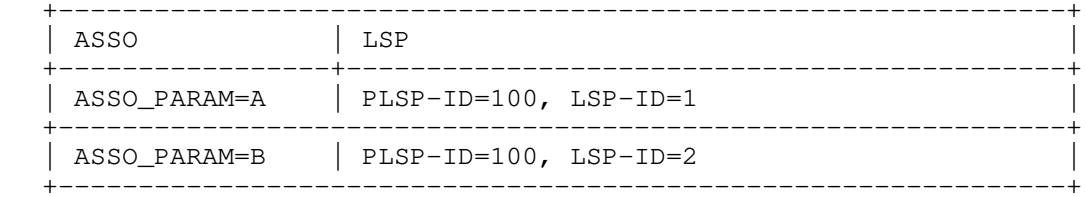


Figure 15: Content of PCE ASSO DB

PCC deletes the old LSP. PCC sends PCRpt(R-FLAG=1, PLSP-ID=100, LSP-ID=1).

ASSO	LSP
ASSO_PARAM=B	PLSP-ID=100, LSP-ID=2

Figure 16: Content of PCE ASSO DB

5. Computation Constraints

For any PCEP object that does not have an explicit removal flag, the absence of that object indicates removal of the constraint specified by that object. For example, suppose the first state-report contains an LSPA object with some affinity constraints. Then if a subsequent state-report does not contain an LSPA object, then this means that the previously specified affinity constraints do not apply anymore. Same applies to all PCEP objects, like METRIC, BANDWIDTH, etc., which do not have an explicit flag for removal. This simply ensures that it is possible to remove a constraint without using an explicit removal flag.

6. Use of RRO, SR-RRO and SRv6-RRO objects

[RFC8231] defines a PCRpt message which contains <intended-path> known as the ERO object and <actual-path> known as the RRO object. [RFC8664] defines SR-ERO and SR-RRO objects for SR-TE LSPs. [I-D.ietf-pce-segment-routing-ipv6] further defines SRv6-ERO and SRv6-RRO objects for SRv6-TE paths.

In practice RRO data set is the result of signalling of the intended path defined in the ERO via protocol such as RSVP. The ERO and RRO values may be different as the path encoded in the ERO may differ than the RRO such as during protection conditions or if the ERO contains loose hops which are expanded upon. As Segment Routing LSP does not perform any signalling, the values of an SR-ERO/SRv6-ERO and SR-RRO/SRv6-RRO (respectively) are in practice the same, therefore some implementations have omitted the SR-RRO/SRv6-RRO when reporting a SR-TE LSP while others continue to send both SR-ERO/SRv6-ERO and SR-RRO/SRv6-RRO values.

A PCC MUST send an (possibly empty) ERO/SR-ERO/SRv6-ERO in the PCRpt message for every LSP. A PCC MAY send an SR-RRO/SRv6-RRO for an SR-TE/SRv6-TE LSP (respectively). A PCE SHOULD interpret the RRO/SR-RRO/SRv6-RRO as the actual path the LSP is taking but MAY interpret only the ERO/SR-ERO/SRv6-ERO as the actual path. In the absence of an RRO/SR-RRO/SRv6-RRO a PCE SHOULD interpret the ERO/SR-ERO/SRv6-ERO (respectively) as the actual path for the LSP.

7. Security Considerations

None at this time.

8. IANA Considerations

None at this time.

9. Acknowledgement

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>.

[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>.

[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>.

[I-D.ietf-pce-segment-routing-ipv6] Li, C., Negi, M., Sivabalan, S., Koldychev, M., Kaladharan, P., and Y. Zhu, "PCEP Extensions for Segment Routing leveraging the IPv6 data plane", draft-ietf-pce-segment-routing-ipv6-08 (work in progress), November 2020.

10.2. Informative References

[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>.

Appendix A. Contributors

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

Email: dhruv.ietf@gmail.com

Andrew Stone
Nokia
Ottawa, Canada

Email: andrew.stone@nokia.com

Mahendra Singh Negi
RtBrick Inc
N-17L, 18th Cross Rd, HSR Layout
Bangalore, Karnataka 560102
India

Email: mahend.ietf@gmail.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

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

Email: pengshuping@huawei.com

Diego Achaval
Nokia

Email: diego.achaval@nokia.com

Hari Kotni
Juniper Networks, Inc

Email: hkotni@juniper.net

PCE Working Group
Internet-Draft
Intended status: Standards Track
Expires: July 19, 2021

B. Rajagopalan
V. Beeram
Juniper Networks
G. Mishra
Verizon Communications Inc.

January 15, 2021

Path Computation Element Protocol(PCEP) Extension for RSVP Color
draft-rajagopalan-pcep-rsvp-color-00

Abstract

This document specifies extensions to Path Computation Element Protocol (PCEP) to carry a newly defined attribute of RSVP LSP called 'color' that can be used as a guiding criterion for selecting the LSP as a next hop for a service route.

Requirements Language

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

Status of This Memo

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

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

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

This Internet-Draft will expire on July 19, 2021.

Copyright Notice

Table of Contents

1. Introduction

This document defines a new RSVP LSP property, called "color", that can be exchanged over PCEP. The 'color' field can be used as one of the guiding criteria in selecting the LSP as a next hop for service prefixes.

While the specific details of how the service prefixes are associated with the appropriate RSVP LSP's are outside the scope of this specification, the envisioned high level usage of the 'color' field is as follows.

The service prefixes are marked with some indication of the type of underlay they need. The underlay LSP's carry corresponding markings, which we refer to as "color" in this specification, enabling an ingress node to associate the service prefixes with the appropriate underlay LSP's.

As an example, for a BGP-based service, the originating PE could attach some community, e.g. the Extended Color Community [RFC5512] with the service route. A receiving PE could use locally configured policies to associate service routes carrying Extended Color Community 'X' with underlay RSVP LSP's of color 'Y'.

While the Extended Color Community provides a convenient method to perform the mapping, the policy on the ingress node is free to

classify on any property of the route to select underlay RSVP LSP's of a certain color.

2. Protocol Operation

The STATEFUL-PCE-CAPABILITY negotiation message is enhanced to carry the color capability, which allows PCC & PCE to determine how incompatibility should be handled, should only one of them support color. An older implementation that does not recognize the new color TLV would ignore it upon receipt. This can sometimes result in undesirable behavior. For example, if PCE passes color to a PCC that does not understand colors, the LSP may not be used as intended. A PCE that clearly knows the PCC's color capability can handle such cases better, and vice versa. Following are the rules for handling mismatch in color capability.

A PCE that has color capability MUST NOT send color TLV to a PCC that does not have color capability. A PCE that does not have color capability can ignore color marking reported by PCC.

When a PCC is interacting with a PCE that does not have color capability, the PCC

- 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" [Value to be assigned by IANA] in LSP State Report message.

If an RSVP tunnel has multiple LSP's associated with it, the PCE should designate one of the LSP's as primary, and attach the color with that LSP. If PCC receives color TLV for an LSP that it treats as secondary, it SHOULD respond with an error code of 4 (Unacceptable Parameters).

3. TLV Format

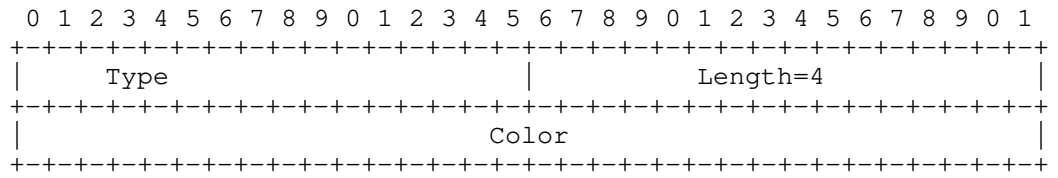


Figure 1: Color TLV in LSP Object

Type has the value [TO-BE-ASSIGNED-BY-IANA]. Length carries a value of 4. The 'color' field is 4-bytes long, and carries the actual color value.

Section 7.1.1 of RFC8231 [RFC8231] defines STATEFUL-PCE-CAPABILITY flags. The following flag is used to indicate if the speaker supports color capability:

C-bit (TO-BE-ASSIGNED-BY-IANA): A PCE/PCC that supports color capability must turn on this bit.

4. Usage with BGP-CT

RSVP LSP's marked with color can also be used for inter-domain service mapping as defined in BGP-CT [I-D.kaliraj-idr-bgp-classful-transport-planes]. In BGP-CT, the mapping community of the service route is used to select a "resolution scheme", which in turn selects LSP's of various "transport classes" in the defined order of preference. The 'color' field defined in this specification could be used to associate the RSVP LSP with a particular transport class.

A colored RSVP LSP can also be exported into BGP-CT for inter-domain classful transport.

5. Security Considerations

This document defines a new TLV for color, and a new flag in capability negotiation, which do not add any new security concerns beyond those discussed in [RFC5440], [RFC8231] and [RFC8281].

An unauthorized PCE may maliciously associate the LSP with an incorrect color. The procedures described in [RFC8253] and [RFC7525] can be used to protect against this attack.

6. IANA Considerations

IANA is requested to assign code points for the following:

- Code point for "Color" TLV from the sub-registry "PCEP TLV Type Indicators".
- C-bit value from the sub-registry "STATEFUL-PCE-CAPABILITY TLV Flag Field".
- An error code for "Unsupported color" from the sub-registry "LSP-ERROR-CODE TLV Error Code Field".

7. Acknowledgments

The authors would like to thank Kaliraj Vairavakkalai, Colby Barth & Natrajan Venkataraman for their review & suggestions, which helped improve this specification.

8. References

8.1. Normative References

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

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

[RFC5512] Mohapatra, P. and E. Rosen, "The BGP Encapsulation Subsequent Address Family Identifier (SAFI) and the BGP Tunnel Encapsulation Attribute", RFC 5512, DOI 10.17487/RFC5512, April 2009, <https://www.rfc-editor.org/info/rfc5512>.

[RFC7525] Sheffer, Y., Holz, R., and P. Saint-Andre, "Recommendations for Secure Use of Transport Layer Security (TLS) and Datagram Transport Layer Security (DTLS)", BCP 195, RFC 7525, DOI 10.17487/RFC7525, May 2015, <https://www.rfc-editor.org/info/rfc7525>.

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

[RFC8253] Lopez, D., Gonzalez de Dios, O., Wu, Q., and D. Dhody, "PCEPS: Usage of TLS to Provide a Secure Transport for the Path Computation Element Communication Protocol (PCEP)", RFC 8253, DOI 10.17487/RFC8253, October 2017, <https://www.rfc-editor.org/info/rfc8253>.

[RFC8281] Crabbe, E., Minei, I., Sivabalan, S., and R. Varga, "Path Computation Element Communication Protocol (PCEP) Extensions for PCE-Initiated LSP Setup in a Stateful PCE Model", RFC 8281, DOI 10.17487/RFC8281, December 2017, <https://www.rfc-editor.org/info/rfc8281>.

8.2. Informative References

[I-D.kaliraj-idr-bgp-classful-transport-planes] Vairavakkalai, K., Venkataraman, N., Rajagopalan, B., Mishra, G., Khaddam, M., and X. Xu, "BGP Classful Transport Planes", draft-kaliraj-idr-bgp-classful-transport-planes-06 (work in progress), January 2021.

Authors' Addresses

Balaji Rajagopalan
Juniper Networks

Email: balajir@juniper.net

Vishnu Pavan Beeram
Juniper Networks

Email: vbeeram@juniper.net

Gyan Mishra
Verizon Communications Inc.

Email: gyan.s.mishra@verizon.com

PCE Working Group
Internet-Draft
Intended status: Standards Track
Expires: August 23, 2021

A. Tokar
S. Sidor
Cisco Systems, Inc.
S. Sivabalan
Ciena
S. Peng
Huawei Technologies
M. Negi
RtBrick Inc

February 19, 2021

Carrying SID Algorithm information in PCE-based Networks. draft-tokar-pce-sid-algo-03

Abstract

The Algorithm associated with a prefix Segment-ID (SID) defines the path computation Algorithm used by Interior Gateway Protocols (IGPs). This information is available to controllers such as the Path Computation Element (PCE) via topology learning. This document proposes an approach for informing headend routers regarding the Algorithm associated with each prefix SID used in PCE-computed paths, as well as signalling a specific SID algorithm as a constraint to the PCE.

Requirements Language

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

Status of This Memo

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

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

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

This Internet-Draft will expire on August 23, 2021.

Table of Contents

1. Introduction

A PCE can compute SR-TE paths using prefix SIDs with different Algorithms depending on the use-case, constraints, etc. While this information is available on the PCE, there is no method of conveying this information to the headend router.

Similarly, the headend can also compute SR-TE paths using different Algorithms, and this information also needs to be conveyed to the PCE for collection or troubleshooting purposes. In addition, in the case of multiple (redundant) PCEs, when the headend receives a path from the primary PCE, it needs to be able to report the complete path information - including the Algorithm - to the backup PCE so that in

HA scenarios, the backup PCE can verify the prefix SIDs appropriately.

An operator may also want to constrain the path computed by the PCE to a specific SID Algorithm, for example, in order to only use SID Algorithms for a low-latency path. A new TLV is introduced for this purpose.

Refer to [RFC8665] and [RFC8667] for details about the prefix SID Algorithm.

This document introduces two new NAI types for the SR-ERO subobject, which is defined in [RFC8664]. A new TLV for signalling SID Algorithm constraint to the PCE is also introduced, to be carried inside the LSPA object, which is defined in [RFC5440].

The mechanisms described in this document are equally applicable to both SR-MPLS and SRv6.

2. Terminology

The following terminologies are used in this document:

ERO: Explicit Route Object

IGP: Interior Gateway Protocol

NAI: Node or Adjacency Identifier.

PCE: Path Computation Element

PCEP: Path Computation Element Protocol.

SID: Segment Identifier.

SR: Segment Routing.

SR-TE: Segment Routing Traffic Engineering.

LSP: Label Switched Path.

LSPA: Label Switched Path Attributes.

3. Object Formats

3.1. SR ERO Subobject

The SR-ERO subobject encoding is extended with additional NAI types.

The following new NAI types (NT) are defined:

- NT=TBD1: The NAI is an IPv4 node ID with Algorithm.
- NT=TBD2: The NAI is an IPv6 node ID with Algorithm.

This document defines the following NAIs:

'IPv4 Node ID with Algorithm' is specified as an IPv4 address and Algorithm identifier. In this case, the NT value is TBD1 and the NAI field length is 8 octets.

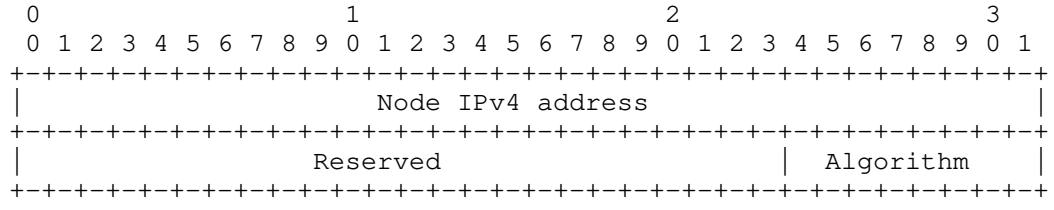


Figure 1: NAI for IPv4 Node SID with Algorithm

'IPv6 Node ID with Algorithm' is specified as an IPv6 address and Algorithm identifier. In this case, the NT value is TBD2 and the NAI field length is 20 octets.

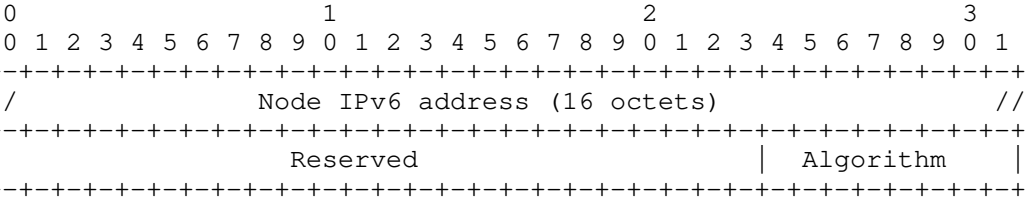


Figure 2: NAI for IPv6 Node SID with Algorithm

3.2. LSPA Object

A new TLV for the LSPA Object with TLV type=TBD3 is introduced to carry the SID Algorithm constraint. This TLV SHOULD only be used when PST (Path Setup type) = SR or SRv6.

The format of the SID Algorithm TLV is as follows:

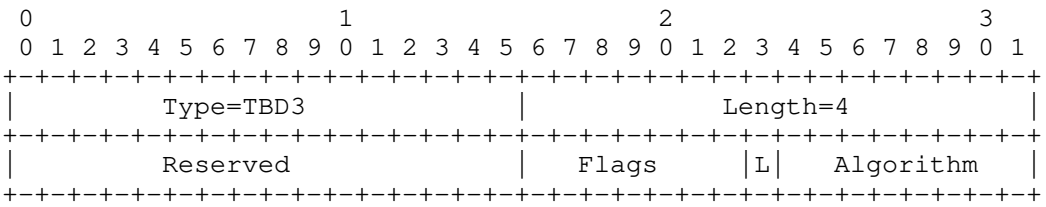


Figure 3: SID Algorithm TLV Format

The code point for the TLV type is TBD3. The TLV length is 4 octets.

The 32-bit value is formatted as follows.

Reserved: MUST be set to zero by the sender and MUST be ignored by the receiver.

Flags: This document defines the following flag bits. The other bits MUST be set to zero by the sender and MUST be ignored by the receiver.

* L (Loose): If set to 1, the PCE MAY insert prefix SIDs with a different Algorithm, but it MUST prefer the specified Algorithm whenever possible.

Algorithm: SID Algorithm the PCE MUST take into acount while computing a path for the LSP.

4. Operation

4.1. SR-ERO NAI Encoding

IPv4 prefix SIDs used by SR-TE paths with an associated Algorithm SHOULD be encoded with 'IPv4 Node ID with Algorithm' NAI.

IPv6 prefix SIDs used by SR-TE paths with an associated Algorithm SHOULD be encoded with 'IPv6 Node ID with Algorithm' NAI.

4.2. SID Algorithm Constraint

In order to signal a specific SID Algorithm constraint to the PCE, the headend MUST encode the SID ALGORITHM TLV inside the LSPA object.

When the PCE receives a SID Algorithm constraint, it MUST only take prefix SIDs with the specified Algorithm into account during path computation. However, if the L flag is set in the SID Algorithm TLV, the PCE MAY insert prefix SIDs with a different Algorithm in order to successfully compute a path.

If the PCE is unable to find a path with the given SID Algorithm constraint, it MUST bring the LSP down.

SID Algorithm does not replace the Objective Function defined in [RFC5541]. The SID Algorithm constraint acts as a filter, restricting which SIDs may be used as a result of the path computation function.

5. Security Considerations

No additional security measure is required.

6. IANA Considerations

6.1. PCEP SR-ERO NAI Types

IANA is requested to allocate new SR-ERO NAI types for the new NAI types specified in this document.

Value	Description	Reference
TBD1	IPv4 Node ID with Algorithm	This document
TBD2	IPv6 Node ID with Algorithm	This document

6.2. PCEP TLV Types

IANA is requested to allocate a new TLV type for the new LSPA TLV specified in this document.

Value	Description	Reference
TBD3	SID Algorithm	This document

7. 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>.

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

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

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

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

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

Appendix A. Contributors

Mike Koldychev
Cisco Systems
Kanata, Ontario
Canada

EMail: mkoldych@cisco.com

Authors' Addresses

Alex Tokar
Cisco Systems, Inc.
Eurovea Central 3.
Pribinova 10
Bratislava 811 09
Slovakia

Email: atokar@cisco.com

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

Email: ssidor@cisco.com

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

Email: msiva282@gmail.com

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

Email: pengshuping@huawei.com

Mahendra Singh Negi
RtBrick Inc
Bangalore, Karnataka
India

Email: mahend.ietf@gmail.com