Path Computation Element Working Group
Internet-Draft
Intended status: Standards Track
Expires: 4 September 2025

0. Dugeon
J. Meuric
Orange Innovation
Y. Lee
Samsung Electronics
D. Ceccarelli
Cisco

3 March 2025

PCEP Extension for Stateful Inter-Domain Tunnels
draft-ietf-pce-stateful-interdomain-07

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. It delivers a new tool in the PCE toolbox in order for operator to build Intent-Based Networking. For this purpose, this document defines a new Stitching Label, new Association Type, new PCEP communication Protocol (PCEP) Capability, new PCE Errors Type and new PCE Notifications Type.

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 and increase drastically the size of the header in the case of SRv6. 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 or remote SRv6 header for an end-to-end inter-domain path.

Finally, Operators would be independant from each other about the technology there are deploying in their respective networks. Thus, the solution should accomodate with various technology and provide a solution to setup inter-domain path compose by a mix of underlay technology i.e. a mix of RSVP-TE and Segment Routing including SR-MPLS and SRv6 dataplane.

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. Binding Label / Segment Identifier (BSID) [RFC9604] 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, with SR-MPLS or SRv6 dataplane, to build inter-domain segment paths. To enable this mechanism, this document defines a new Stitching Label, new Association Type, new PCEP communication Protocol (PCEP) Capability, new PCE Errors Type and new PCE Notifications Type.

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.

* Each PCC is able to configure a Binding Label/Segment Identifier (BSID) and each PCE is able to request a BSID to a PCC or a neighbor domains' PCE.

* 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.

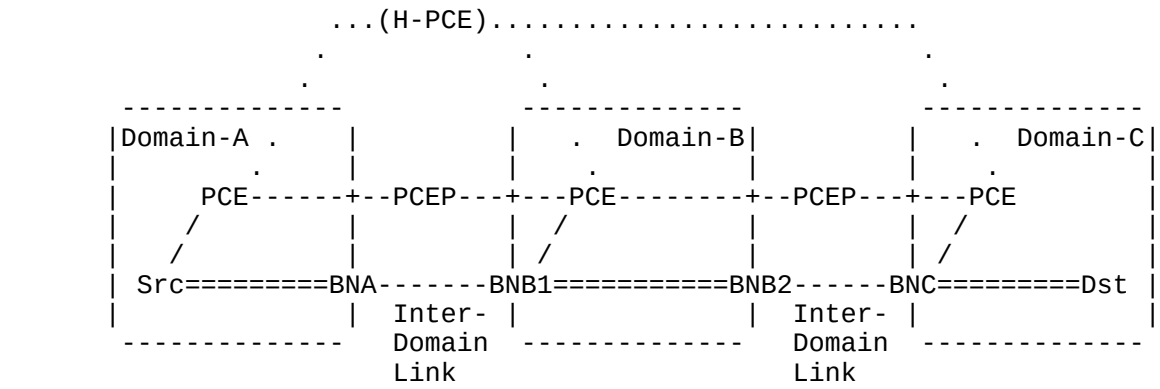

Figure 1: Example of the representation of 3 domains with 3 PCEs

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.

BSID: Binding Label / Segment Identifier.

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).

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) and is conveyed as an inter-domain BSID.

TPB(): An empty TE-PATH-BINDING TLV to request an inter-domain BSID i.e. a Stitching Label.

TPB(i): A TE-PATH-BINDING TLV with an inter-domain Binding Value equal to the Stitching Label SL(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) or SR Policy could be used for that purpose. But, when stitching or nesting tunnels, the FEC or SR Policy 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 for RSVP-TE and SR -MPLS and to refer to SRv6 SID for SRv6 Dataplane. 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. Binding Label / Segment Identifier (BSID) [RFC9604] defines a new TE-PATH-BINDING TLV to exchange a Binding Segment / Label Identifier (BSID) between a PCC and a PCE. This BSID is then used to steer incoming traffic using this BSID into the associated path. Thus, the Stitching Label defines in this draft is a particular use case of BSID, named inter-domain BSID, and could be conveyed in the TE-PATH-BINDING TLV of the LSP Object without any modification of PCEP nor PCEP Objects.

2.2. Inter-domain traffic steering

If BSID allows to automatically steer traffic identified with this BSID into the associated path, for inter-domain BSID, it is different as the Stitching Label is associated to the inter-domain link LK(i+1) i.e. the link between the border node BN-ex(i) of the domain(i) and the border node BN-en(i+1) of the domain(i+1). Indeed, the Border Node BN-en(i+1) needs to received the traffic identified by the Stitching Label SL(i+1) from BN-ex(i). Thus, it is necessary to instruct the border node BN-ex(i) to push the Stitching Label(i+1) on top of the packets of traffic going from domain(i) to domain(i+1), and send them to the border node BN-en(i+1) through the inter-domain link LK(i+1). Depending of technology used by domain(i), RSVP-TE or Segment Routing, the operation uses two different approaches.

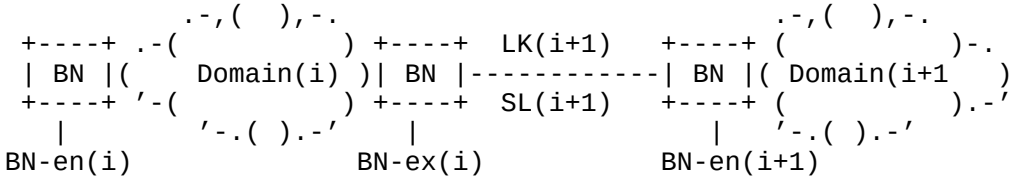

Figure 2: Inter-domain Link

2.2.1. Stitching RSVP-TE

In case of RSVP-TE, the Border Node BN-ex(i) needs to receive 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.

Thus, PCE(i) SHOULD provide SL(i+1) and LK(i+1) to the PCC BN-en(i) through 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). It is left to implementation of PCE to get the LK(i+1) value. One solution consist to retrieve it from the PKS(i) or the ERO previously computed during the BRPC process.

2.2.2. Stitching Segment Routing

In case 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 Binding SID entry for BN-en(i+1) to identify the packets that follow the next Segment Path. For that purpose, BN-en(i) MUST install in its MPLS L(F)IB an instruction to replace the incoming Stitching Label SL(i) by the label stack given by the ERO(i) plus the Stitching Label SL(i+1).

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 [RFC9086] for that purpose. The inter-domain SID named here I-SID(i+1) is announced by BN-ex(i) to PCE(i) through BGP-LS for each interface that connects BN-ex(i) to neighbors BN-en(i+1). Thus, PCE(i) SHOULD provide SL(i+1) and I-SID(i+1) to the PCC BN-en(i) through the ERO, so that the label stack will end with {BN-ex(i) SID, I-SID(i+1), 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.

2.2.3. Stitching SRv6

Segment Routing over IPv6 (SRv6) Network Programming [RFC8986] offers the possibility to specify a packet processing program by encoding a sequence of instructions in the IPv6 packet header. Binding Label / Segment Identifier (BSID) [RFC9604] offers also the possibility to request an SRv6 Binding SID to the PCC. Thus, similar to SR-MPLS case, PCE will compute the SRv6 Policy and insert the Stitching Label SL(i+1) represented by an SRv6 SID as the last SRv6 Locator in the SRv6 Header (SRH). Thus, when the packet reach the BN-en(i+1) router, the Stitching Label SL(i+1) serves as an SRv6 entry to identify the next Segment Path. For that purpose, the requested SRv6 Binding SID MUST be composed of an SRv6 SID with Behavior and Structure in order for the BN-en(i+1) to determine how to route packet to the next Segment Path. For that purpose, End.B6 or End.B6.Encaps MUST be used.

Like for SR-MPLS case, when there are multiple interfaces between Border Nodes, BN-ex(i) needs to know how to send the packets to BN-en(i+1). Similary to SR-MPLS and Egress Peer Engineering, it is RECOMMENDED to use BGP-LS extension for SRv6 [RFC9514] to collect the different SRv6 BGP Peer Node SID. Then, PCE MUST select the appropriate SRv6 BGP Peer SID during the path computation in particular for BN-ex(i) to send the traffic to BN-en(i+1) on the appropriate interface. This time, an SRv6 End.X MUST be used and inserted before the Stitching SRv6 SID. The global operation is as follow:

* The penultimate router of domain(i) selects the next SRv6 Locator in the Segment Routing Header (SRH) and sends the packet to the next node designated by this SRv6 SID which corresponds to router BN-ex(i).

* Then, BN-ex(i) selects the next SRv6 Locator in the SRH i.e. a SRv6 End.X SID which corresponds to the selected Adjacency link between the router BN-ex(i) and the router BN-en(i+1) and sends packets to BN-en(i+1) on the appropriate link.

* In turn, BN-en(i+1) selects the next SRv6 Locator in the SRH which is the last one in the stack and corresponds to the Stitching Label SL(i+1). This SRv6 SID End.B6 or End.B6.Encaps instructs the router BN-en(i+1) to pop this SRv6 header and to replace it by the new SRH that corresponds to this Binding SID. Packets are then forwarded in domain i+1 according to the new SRv6 header.

2.2.4. Strict traffic steering

The Binding Label / Segment Identifier has been defined as a global traffic steering identifier. Thus, if an entry border node BN-en(i) is configured with a Stitching Label SL(i), any domain connected to this border node through different interface could send traffic to domain(i) and subsequent domains even if they are not part of the inter-domain path. However, some operators would prefer to configure a strict enforcement of traffic steering. In this case, the border node BN-en(i) SHOULD restrict the MPLS L(F)IB configuration or SRv6 Endpoint Behavior to accept traffic with the Stitching Label SL(i) from only the incoming link LK(i).

2.3. Inter-domain Flags for TE-PATH-BINDING TLV

In order to convey the Stitching Label and manage traffic steering at inter-domain, this specification defines new flags (See IANA section of this document) for the Binding Label / Segment Identifier. The format of the TE-PATH-BINDING TLV is defined in Binding Label / Segment Identifier (BSID) [RFC9604] and included here for easy reference with the addition of the new flags as follow:

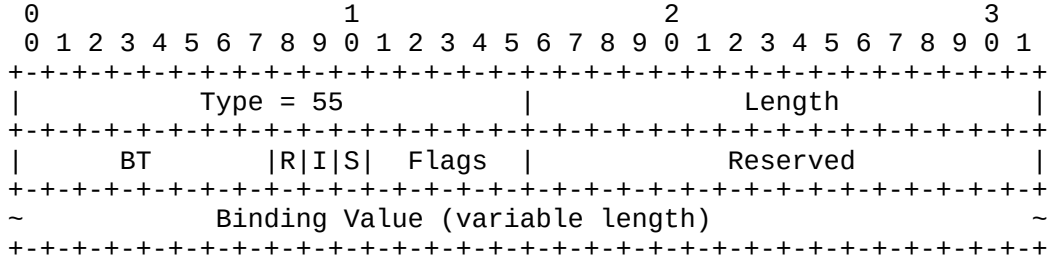

Figure 3: TE-PATH-BINDING TLV

* I flag: Inter-Domain Binding indicates that this Binding Value corresponds to an inter-domain path, thus that this Binding Value is a Stitching Label.

* S flag: Strict Binding indicates that the PCC MUST restrict the Binding Value to the interface that corresponds to the domain source End-Point of the associated path and MUST reject incoming traffic with this Binding Value when it reaches the PCC through another interface.

2.4. Operations

An empty TE-PATH-BINDING TLV with the I flag set to 1 MUST be present in a PCInitiate messages sent 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(i) or Child PCE(i) MUST return a Stitching Label SL in the TE-PATH-BINDING TLV with the I flag set in the LSP object of the PCRpt message to PCE(i-1) or the Parent PCE. PCE(i-1) MUST NOT provide a Stitching Label as a Binding Value of the TE-PATH-BINDING TLV to its neighbor PCE(i).

An empty TE-PATH-BINDING TLV with the I flag set to 1 MUST be present in the PCInitiate message sent 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. The I flag MUST be set by the PCE(i) only after receiving a PCInitiate message with an empty TE-PATH-BINDING with the I flag set 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 in the TE-PATH-BINDING TLV with the I flag set in the LSP object of the PCRpt message. Alternatively, the PCE(i) could provide a Stitching Label as a Binding Value of the TE-PATH-BINDING TLV with the I flag set to the PCC of the domain(i) when initiating a new local path(i) as per section #8 of draft Binding Label / Segment Identifier (BSID) [RFC9604]. If the PCC is not able to allocate a BSID for inter-domain, it MUST send a PCErr message with Error-Type = 32, "Binding label/SID failure" and Error-Value = 2, "Unable to allocate a new binding label/SID" defined in draft Binding Label / Segment Identifier (BSID) [RFC9604].

If a PCE(i) receives a PCRpt without a TE-PATH-BENDING TLV while it has requested a Stitching Label in the PCInitiate message, it MUST send a PCErr message with Error-Type = 6, "Mandatory Object missing"" and Error-Value = TBD1, "LSP TE-PATH-BINDING missing TLV". If a PCE(i) receives a PCRpt with a TE-PATH-BENDING TLV with the I flag unset while it has requested a Stitching Label in the PCInitiate message, it MUST send a PCErr message with Error-Type = 32: "Binding label/SID failure" and Error-Value = 3: "Unable to allocate a new binding label/SID".

PCE(i) SHOULD set the S flag in addition to the I flag if it requests traffic steering strictly coming from a given interface, i.e. traffic using the BSID and coming from a different interface MUST be rejected by the PCC. When the S flag is set, PCE(i) MUST set the EndPoint source address of the requested local path with the IP address of the interface where the traffic is strictly steered. When the PCC receives an LSP object with an empty TE-PATH-BINDING TLV and

the S flag set, it MUST allocate a Binding Value and configure its MPLS L(F)IB to accept traffic with this BSID only coming from the interface identified by the source address of the EndPoint Object. If the PCC is not be able to strictly steer traffic, it MUST send a PCErr message with Error-Type = "Binding label/SID failure" and Error-Value = "Unable to allocate a new binding label/SID".

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 described in RFC5441 [RFC5441]. Note that for all describe processes in this section, the Sitching Label could corresponds to an MPLS label for RSVP-TE, an MPLS label or Index for Segment Routing or an SRv6 Segment Identifier with endpoint behavior.

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 be able to determine the next 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}, an LSP Object containing an empty TE-PATH-BINDING TLV with the I flag set and the 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 an LSP containing an empty TE-PATH-BINDING TLV with I flag set and ERO = {PKS(i), PKS(i+1), ..., PKS(n), D)}, it MUST sends the received PCInitiate message to PCE(i+1) with a popped ERO and records its received PKS(i) part. All PCE(i)s MUST 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 detailed ERO(n) from the PKS(n) if necessary and MUST send to BN-en(n) a PCInitiate message with the ERO(n) = {BN-en(n), ..., D}, an LSP containing an empty TE-PATH-BINDING TLV with the I flag set 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 service and that it MUST allocate a Binding Value for this 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 or Segment Routing, accordingly to the PST value, 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 MUST send a PCRpt message to its PCE(n) including PLSP-ID(n) and a TE-PATH-BINDING TLV with the Binding Value equal to SL(n) and the I flag set

4. Once PCE(n) receives the PCRpt message from the PCC BN-en(n) with the RRO, PLSP-ID and TE-PATH-BINDING TLV with the I flag set, it MUST send to the PCE(n-1) a PCRpt message containing the TE-PATH-BINDING TLV it received from the PCC BN-en(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 an LSP object containing PLSP-ID(i+1) and a Binding Value in the TE-PATH-BINDIG TLV with the I flag set, it retrieves the detailed ERO(i) from the PKS(i), recorded in step 1, if necessary. Then, it MUST send to the PCC BN-en(i) a PCInitiate message with this ERO(i), an LSP object containing an empty TE-PATH-BINDING TLV with the I flag set and the 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 and that it MUST allocate a Binding Value for this path. PCE(i) sets Path Setup Type (PST) to 0, respectively to 1 to instruct the PCC to enforce the local path by means of RSVP-TE respectively Segment Routing.

2. Egress Control mechanism, as per RFC4003 section 2.1 [RFC4003] for RSVP-TE, respectively, Egress Peer Engineering [RFC9086] for Segment Routing, is used to stitch and steer traffic between the border node BN-ex(i) and BN-en(i+1). This allow PCE(i) to instruct the egress node of domain(i), i.e. BN-ex(i), to forward packets belonging to this tunnel with the Stitching Label. For

that purpose, PCE(i) should identify the link LK(i+1) by retrieving from the PKS(i) the corresponding IP address of the link LK(i+1) for RSVP-TE or from the BGP-LS the label that could be used to reach link LK(i+1) for Segment Routing. 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). Thus, PCE(i) MUST complete the ERO(i), in order to provide the Stitching Label SL(i+1) and Link identifier LK(i+1) to the PCC, as the last hop of the local path i.e. ERO(i) = {ERO(i), [LK(i+1), SL(i+1)]}.

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 or Segment Routing, accordingly to the PST value, with the given ERO(i).

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 MUST send a PCRpt message to its PCE(i) including PLSP-ID(i) and a TE-PATH-BINDING TLV with the I flag set containing a Binding Value equal to SL(i).

5. Once PCE(i) receives the PCRpt message from the PCC BN-en(i) with the RRO PLSP-ID and TE-PATH-BINDING TLV with the I flag set, it MUST send to the PCE(i-1) a PCRpt message containing the TE-PATH-BINDING TLV it received from the PCC BN-en(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 TE-PATH-BINDING TLV with the I flag set containing the Binding Value equal to the Stitching Label SL(2), it MUST send a PCInitiate message to PCC node S with ERO equal to {ERO(1), [LK(2), SL(2)]}, once retrieves the identifier of link LK(2), and End-Points Object = {S, BN-ex(1)}. This time, no TE-PATH-BINDING TLV is provided 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)

* TPB(I) = Empty TE-PATH-BINDING TLV with the I flag set

* TPB(I, SL) = TE-PATH-BINDING TLV with Binding Value equal to Stitching Label SL and the I flag set

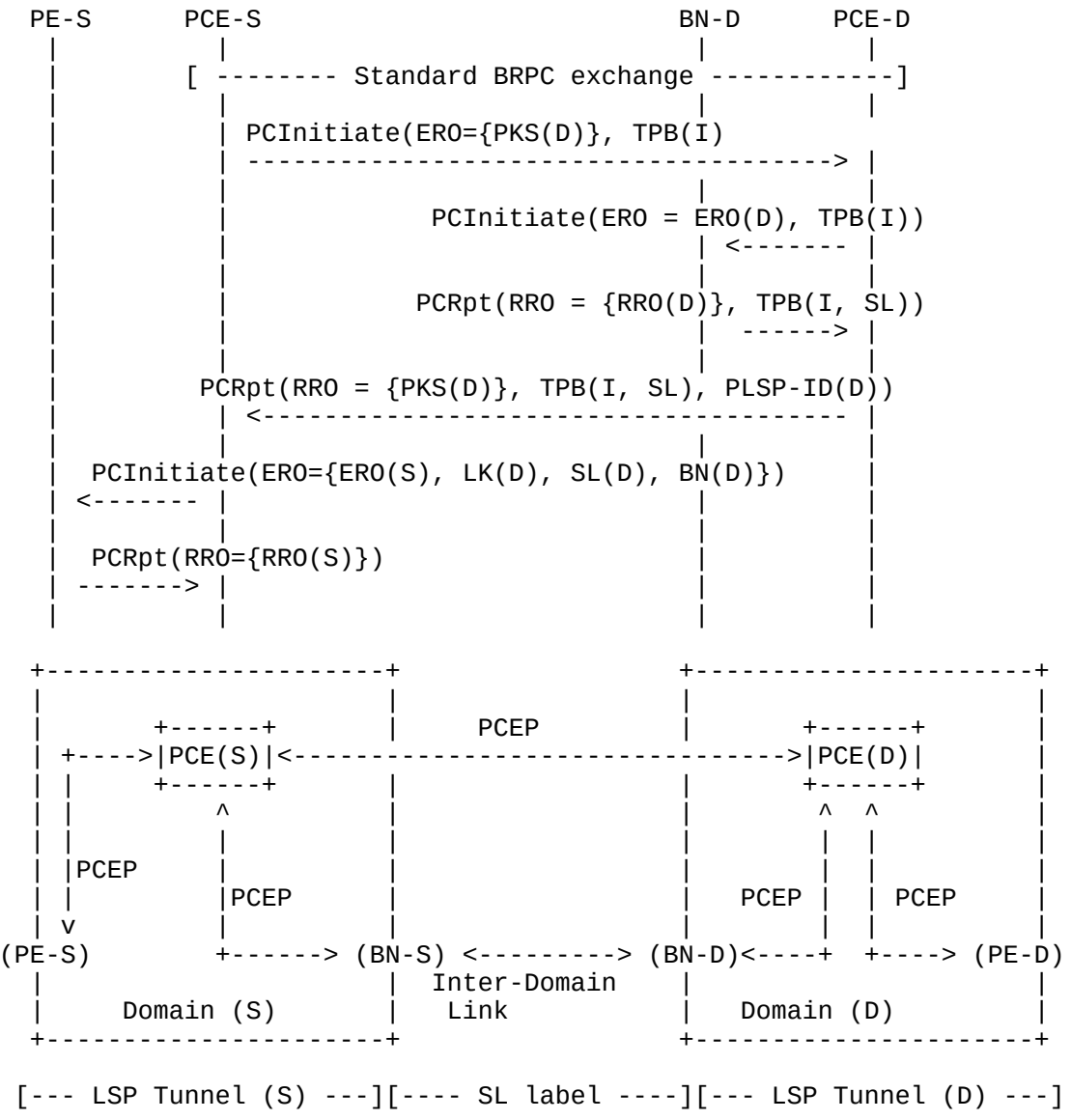

Figure 4: Example of inter-domain path setup between two domains

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 I flag of the TE-PATH-BINDING TLV defined in this document is not supported by the neighbor PCE or PCC, the PCE, respectively PCC, MUST return a PCErr message with Error-Type = "Binding label/SID failure" and Error-Value = "Unable to allocate a new binding label/SID" (as per section #12 of draft Binding Label / Segment Identifier (BSID) [RFC9604]) to its neighbor PCE respectively PCE.

If a PCE(i) receives a PCInitiate message from its peer PCE(i-1) without an TE-PATH-BINDING with the I flag set in the LSP object, it MUST return a PCErr message with Error-Type = 24 (LSP instantiation error) and Error-Value = 1 (Unacceptable instantiation parameters) to its peer PCE(i-1).

Following a PCInitiate message with an LSP object containing an empty TE-PATH-BINDING TLV with the I flag set, if a neighbor PCE(i+1) or a PCC returns no TE-PATH-BINDING TLV, or a TE-PATH-BINDING TLV without the I flag set, the PCE(i), respectively the PCE(i), MUST return a PCErr message with Error-Type = "Binding label/SID failure" and Error-Value = "Unable to allocate a new binding label/SID".

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]. Note that for all describe processes in this section, the Sitching Label could corresponds to an MPLS label for RSVP-TE, an MPLS label or Index for Segment Routing or an SRv6 Segment Identifier with endpoint behavior.

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 follows the different steps described below:

Step 1: Initialization

1. The Parent PCE MUST send a PCInitiate message to Child PCE(n) with an ERO = {PKS(n)} an LSP containing an empty TE-PATH-BINDING TLV with the I flag set and End-Points = {BN-en(n), D}. Then, PCE(n) retrieves the ERO from the PKS(n), if necessary, and MUST send to BN-en(n) a PCInitiate message with the ERO(n) = {BN-en(n), ..., D}, an LSP Object with empty TE-PATH-BINDING TLV with the I flag set 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 and that it MUST allocate a Binding Value for this 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 or Segment Routing, accordingly to the PST value, 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 MUST send a PCRpt message to its PCE(n) with PLSP-ID(n) and a TE-PATH-BINDING TLV with the I flag set and a Binding Value equal to SL(n).

4. Once PCE(n) receives the PCRpt message from the PCC BN-en(n) with the RRO, PLSP-ID and TE-PATH-BINDING TLV with the I flag set, it MUST send to its Parent PCE a PCRpt message containing the TE-PATH-BINDING TLV it received from the PCC BN-en(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. Once the Parent PCE receives a PCRpt message from Child PCE(i+1), it MUST send a PCInitiate message to Child PCE(i) with an LSP object containing an empty TE-PATH-BINDING TLV with the I flag set, the ERO(i) to which it appends the SL(i+1) i.e. ERO(i) = {PKS(i), SL(i+1)} and End-Points = {BN-en(i), BN-ex(i)}.

2. When PCE(i) receives a PCInitiate message from its Parent PCE, it retrieves the detailed ERO(i) from the PKS(i) if necessary. Then, it MUST send to the PCC BN-en(i) a PCInitiate message with an LSP object containing an empty TE-PATH-BINDIG TLV with the I flag set, this ERO(i) 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 and that it MUST allocate a Binding Value for this path. PCE(i) sets Path Setup Type (PST) to 0, respectively to 1 to instruct the PCC to enforce the local path by means of RSVP-TE respectively Segment Routing.

3. Egress Control mechanism, as per RFC4003 section 2.1 [RFC4003] for RSVP-TE, respectively, Egress Peer Engineering [RFC9086] for Segment Routing, is used to stitch and steer traffic between the border node BN-ex(i) and BN-en(i+1). This allow PCE(i) to instruct the egress node of domain(i), i.e. BN-ex(i), to forward packets belonging to this tunnel with the Stitching Label. For that purpose, PCE(i) should identify the link LK(i+1) by retrieving from the PKS(i) the corresponding IP address of the link LK(i+1) for RSVP-TE or from the BGP-LS the label that could be use to reach link LK(i+1) for Segment Routing. 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). Thus, PCE(i) MUST complete the ERO(i), in order to provide the Stitching Label SL(i+1) and Link identifier LK(i+1) to the PCC, as the last hop of the local path i.e. ERO(i) = {ERO(i), [LK(i+1), SL(i+1)]}.

4. 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 or Segment Routing, accordingly to the PST value, with the given ERO(i).

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 MUST send a PCRpt message to its PCE(i) with PLSP-ID(i) and a TE-PATH-BINDING TLV with I flag set and a Binding Value equal to SL(i).

6. Once PCE(i) receives the PCRpt message from the PCC BN-en(i) with the RRO, PLSP-ID and TE-PATH-BINDING TLV with the I flag set, it MUST send to its Parent PCE a PCRpt message containing the TE-PATH-BINDING TLV it received from the PCC BN-en(i) and the PLSP-ID(i). PCE(i) MAY add {PKS(i), ..., PKS(n)} in the RRO.

7. Once the Parent PCE receives the PCRpt message 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 MUST send a last PCInitiate message to its Child PCE(1) with an LSP Object containing an empty TE-PATH-BINDING TLV with the I flag set, ERO = {PKS(1), SL(2)} and End-Points = {S, BN-ex(1)}. In turn, Child PCE(1) MUST send a PCInitiate message to PCC node S with ERO equal to {ERO(1), [LK(2), SL(2)]} and End-Points Object = {S, BN-ex(1)}. This time, no TE-PATH-BINDING TLV is provided 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.

4.2. Completion Failure of Inter-domain Path Setup Procedure

In case of error during path set up, PCRpt and/or PCErr messages MUST be used to signal the problem to the Parent PCE. In particular, if the new I flag of the TE-PATH-BINDING TLV defined in this document is not supported by the Child PCE or the PCC, the Child PCE, respectively the PCC, MUST return a PCErr message with Error-Type = "Binding label/SID failure" and Error-Value = "Unable to allocate a new binding label/SID" (as per section #12 of draft Binding Label / Segment Identifier (BSID) [RFC9604]) to its Parent PCE respectively PCE.

If a PCE(i) receives a PCInitiate message from its Parent PCE without a TE-PATH-BINDING with the I flag set in the LSP, it MUST return a PCErr message with Error-Type = 24 (LSP instantiation error) and Error-Value = 1 (Unacceptable instantiation parameters) to its Parent PCE.

Following a PCInitiate message with an LSP containing an empty TE-PATH-BINDING TLV with the I flag set, if a Child PCE or a PCC returns no TE-PATH-BINDING TLV, or a TE-PATH-BINDING TLV without the I flag set, the Parent PCE, respectively the Child PCE, MUST return a PCErr message with Error-Type = "Binding label/SID failure" and Error-Value = "Unable to allocate a new binding label/SID".

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.

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.

4.3. Example for Stateful H-PCE Stitching Procedure

Taking the sample hierarchical domain topology example from [RFC6805] as the reference topology for the entirety of this section.

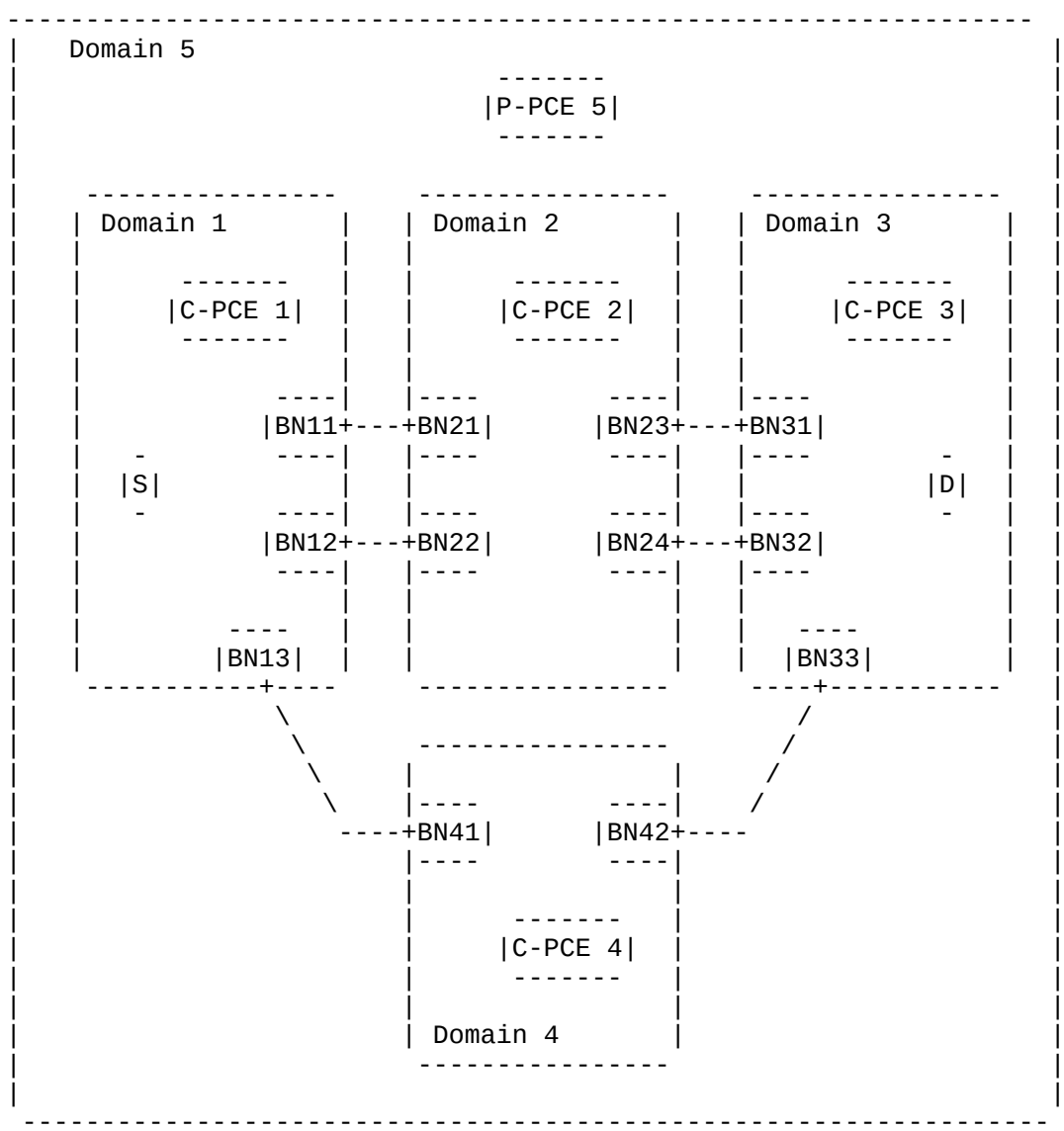

Figure 5: Hierarchical domain topology from RFC6805

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 an LSP object containing an empty TE-PATH-BINDING TLV with the I flag set and PST = 0/1 based on the setup type.

2. C-PCE3 further propagates the initiate message it receives from P-PCE to BN33.

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) as the Binding Value of the TE-PATH-BINDING TLV with the I flag set and the RRO through the PCRpt message.

6. C-PCE3 further reports the PCRpt message it receives from BN33 to the P-PCE (P-PCE5).

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,SL33,BN33} and an LSP object containing an empty TE-PATH-BINDING TLV with the I flag set and PST = 0/1 based on the setup type.

2. C-PCE4 further propagates the initiate message it receives from P-PCE to BN41 once complete the the ERO with the Link Identifier LK33 i.e. ERO={BN41..BN42,LK33,SL33,BN33}.

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) as the Binding Value of the TE-PATH-BINDING TLV with the I flag set and the RRO through the PCRpt message.

6. C-PCE4 further reports the PCRpt message it receives from BN41 to the P-PCE (P-PCE5).

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,SL41,BN41} and an LSP object containing an empty TE-PATH-BINDING TLV with the I flag set and PST = 0/1 based on the setup type.

2. C-PCE1 further propagates the initiate message it receives from P-PCE to node S once complete the the ERO with the Link Identifier LK41 i.e. ERO={S..BN13,LK41,SL41,BN41}.

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 PCRpt message it receives from node S to the P-PCE (P-PCE5).

5. Inter-domain Path Management

This section describes how inter-domain paths could be managed.

5.1. Inter-domain PCE Capabilities

A PCE needs to know if its neighbor PCEs as well as PCCs are able to participate and setup an inter-domain path. This document defines new PCEP Capability for the recursive method and defines new flag for the hierarchical method.

5.1.1. Inter-domain PCE Capability

A new PCE Capabilities is defined for the recursive method and the capabilities to setup inter-domain path. The INTER-DOMAIN-PCE-CAPABILITY TLV is an optional TLV for use in the OPEN object for PCE capability advertisement. Its format is shown in the following figure:

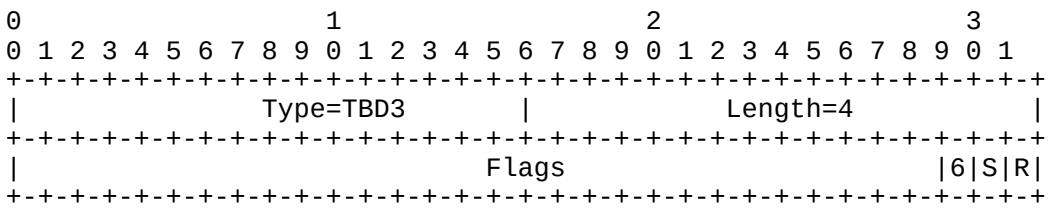

Figure 6: INTER-DOMAIN-PCE-CAPABILITY TLV Format

The Type (16 bits) of the TLV is TBD3. The Length field is 16 bits long and has a fixed value of 4.

The value comprises a single 32 bits "Flags" field:

* R (Recursive Path Computation - 1 bit): the R flag indicates that the PCE is supporting a recursive path computation as per BRPC [RFC5441] in order to compute an end-to-end path. This flag MUST be set only by a PCE when establish a PCEP session with a peer PCE. PCC MUST keep this flag unset.

* S (INTER-DOMAIN-STITCHING-LABEL-CAPABILITY - 1 bit): if set to 1 by a PCE, the S flag indicates that the domain is supporting Stitching Label to set up inter-domain paths i.e. the PCE is able to participate to an inter-domain path setup by requesting to its PCC and propagating to its neighbors PCE a Stitching Label. When set by a PCC, the S flag indicates that the PCC is able to provide a Stitching Label as value of TE-PATH-BINDING TLV when a PCE request an Inter-Domain Binding SID.

* 6 (INTER-DOMAIN-STITCHING-SRV6-CAPABILITY - 1 bit): if set to 1 by a PCE, the 6 flag indicates that the domain is supporting Stitching by means of SRv6 to setup inter-domain paths i.e. the PCE is able to participat to an inter-domain path setup by requesting to its PCC and propagating to its neighbors PCE an SRv6 BSID as value of TA-PATH-BINDING TLV when a PCE request an Inter-Domain Binding SID.

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

A PCE MUST set the R flag when establishing a PCEP session with a neighbor PCE when performing recursive end-to-end path computation when adding Inter-Domain Capability to the PCEP Open Message.

A PCE MUST set the S and/or 6 flag when establishing a PCEP session with a neighbor PCE when adding Inter-domain Capability to the PCEP Open Message. Both S and 6 flags MUST be set when PCE is able to perform interworking between SRv6 and MPLS.

A PCC MUST set the S or 6 flag when adding the Inter-Domain Capability to the PCEP Open Message when establishing a PCEP session with a PCE.

A PCC MUST leave the R flags unset when adding Inter-Domain Capability to the PCEP Open Message when establishing a PCEP session with a PCE. If a PCE receives an Inter-Domain Capability within the PCEP Open message with R flag set from a PCC, then the session establishment MUST fail, and the PCE MUST respond with a PCErr message using Error-Type 1 (PCEP session establishment failure) and Error-Value 3 (unacceptable and non negotiable session characteristics).

5.1.2. Hierarchical Inter-domain Capability

In order to determine if a PCE supports the Stitching Label capability, this specification defines new flag (See IANA section of this document) for the H-PCE-CAPABILITY TLV. The format of the PCEP H-PCE-CAPABILITY TLV is defined in H-PCE [RFC8685] and included here for easy reference with the addition of the new flag as follow:

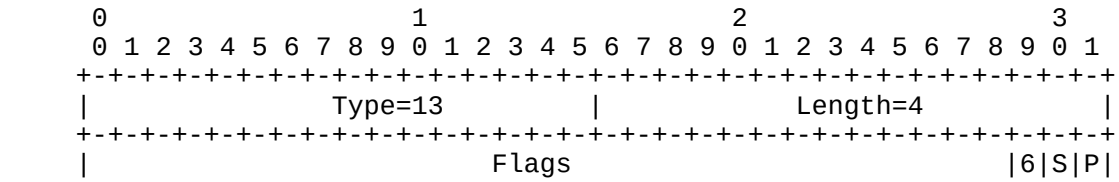

Figure 7: H-PCE-CAPABILITY TLV Format

* S (INTER-DOMAIN-STITCHING-LABEL-CAPABILITY - 1 bit): if set to 1 by a PCE, the S flag indicates that the domain is supporting Stitching Label to set up inter-domain paths i.e. the PCE is able to participate to an inter-domain path setup by requesting to its PCC and propagating to its neighbors PCE a Stitching Label. When set by a PCC, the S flag indicates that the PCC is able to provide a Stitching Label as value of TE-PATH-BINDING TLV when a PCE request an Inter-Domain Binding SID.

* 6 (INTER-DOMAIN-STITCHING-SRV6-CAPABILITY - 1 bit): if set to 1 by a PCE, the 6 flag indicates that the domain is supporting Stitching by means of SRv6 to setup inter-domain paths i.e. the PCE is able to participat to an inter-domain path setup by requesting to its PCC and propagating to its neighbors PCE an SRv6 BSID as value of TA-PATH-BINDING TLV when a PCE request an Inter-Domain Binding SID.

A PCE MUST set the S and/or 6 flag when establishing a PCEP session with a neighbor PCE to signal that is able to support the inter-domain stateful capability within the H-PCE-CAPABILITY TLV to the PCEP Open Message. Both S and 6 flags MUST be set when PCE is able to perform interworking between SRv6 and MPLS.

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

After a failure or reboot, when PCE(i) starts, it will receive PCRpt messages from its PCCs and neighbors PCE(i+1) to synchronize the Inter-domain paths. In addition, it may receive PCInitiate messages from its previous neighbors PCE(i-1) to re-initiate its inter-domain path part. As the PCE(i) may lost the PLSP-ID association, or PLSP-ID change, a new association group (within Association Object) is used to ease the association of the distinct 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 within the PCEP ASSOCIATION OBJECT defined in PCEP Extensions for Establishing Relationships between Sets of Label Switched Paths [RFC8697], and included here for easy reference, is defined with the addition of the new Type value TBD4 as follow:

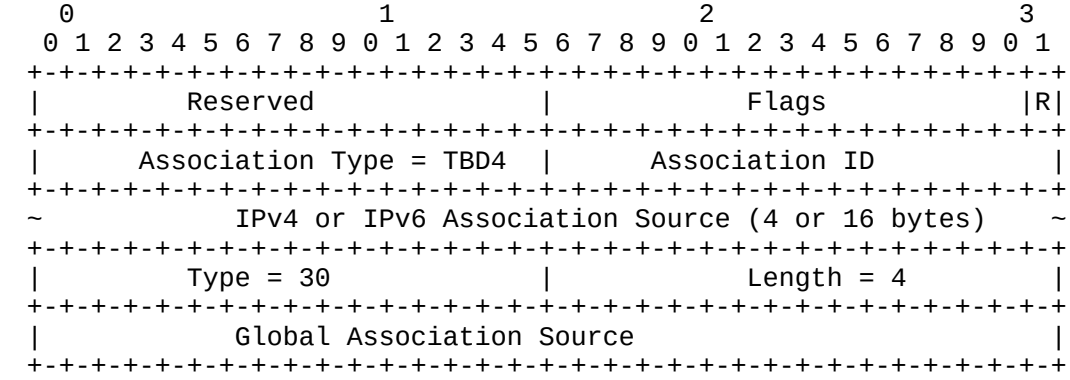

Figure 8: New Inter-Domain Association Group

* 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.

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:

Subsequent PCE(i), for i = 2 to n, MUST send this Association Object unmodified 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 TBD2 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.

5.4. Modification of Inter-Domain Paths

The modification of the inter-domain path is more complex as for a single domain. Indeed, it implies the similar coordination of the PCE(i) along the inter-domain path. Two scenarios need also to be distinguish which have not the same impact:

* BN-en(1..n) and BN-ex(1..n) nodes are not modified: This is the simple case where a simple coordination with PCUpd message could be achieved following the same principle as for the PCInitiate message,

* At least one of BN-en(1..n) or BN-ex(1..n) nodes is modified: This is the general case which imply a complex coordination between PCE(i).

In addition, only PCE(1), respectively the Parent PCE, as PCEP initiator of the inter-domain path is able to modify the inter-domain path and thus modify the BN-en(1..n) and BN-ex(1..n) nodes. However, if a PCE(i) needs to modify its local path which could imply a modification of the BN-en(i) and/or the BN-ex(i) node, it could send

back a PCRpt message to PCE(1) with a LSP Object including Operational bits set to "Going-Down" i.e. value 3. But, there is no guarantee that the PCE(1) or parent PCE will trigger a new inter-domain path computation to update the inter-domain path when it receives such report.

Thus, when modification of inter-domain path implies the modification of a border node between two domains, the operation MUST follow the "Make-Before-Break" principle. Indeed, the coordination between the various PCE(i) will take time and all domains involved in the chain must be sure that the new path is in place before switching the on-going traffic to the new path to avoid packets loss. Border nodes modification implies that, at least, one of the source or destination of a local path that form the inter-domain path has been modified. Thus, PCUpd message could not be used as source or destination of the tunnel has been altered. A PCInitiate message must be used instead to setup the new local tunnel. However, once the new local path is in place between domain (i) and domain (i+1), PCE(i), respectively PCE(i+1) MUST be advertise that the operation is complete to i) switch traffic from old local path to new local path and ii) remove their respective old local paths. This is especially true when a domain withdrawn from the inter-domain path and replace by a new one: it must be advertised to remove the previously allocated resources.

For that purpose, a new PCEP Notification Object has been defined in order to manage complex inter-domain path update and to give the possibility of intermediate domain to trigger a modification.

5.4.1. Inter-Domain PCEP Notification

To help the inter-domain path management operations, a new Notification Type of the PCNtf message with new Notification Value is proposed to advertise the PCE(1) or parent PCE that the inter-domain path needs an update and for PCE(1) or parent PCE to advertise PCE(i) that the update process has been completed.

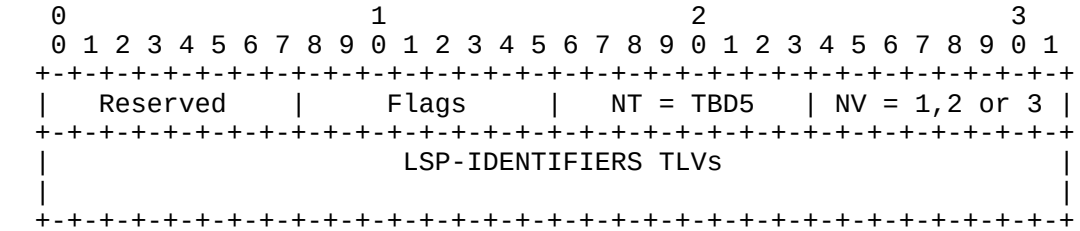

Figure 9: New PCEP Notification Type and Value with LSP-ID TLV

* Notification-Type = TBD5: Inter-domain path needs attention

* - Notification-Value = 1: PCE(i) requests an inter-domain path update to the PCE(1)

  - Notification-Value = 2: PCE(i) advertises PCE(1) that it could no longer participate to this inter-domain path

  - Notification-Value = 3: PCE(1) advertises PCE(2..n) that the inter-domain path has been updated

* LSP-IDENTIFIERS TLVs MUST be present and set to the PLSP-ID(i) of the inter-domain path advertised by the PCE(i) to the PCE(i-1).

When a PCE(i) needs to modify its local part of the inter-domain path which imply the modification of the BN-en(i) and/or the BN-ex(i) node, it MUST send a PCNtf message backward to PCE(1) for the Backward Recursive method or the parent PCE for the Hierarchical method with Notification-Type = TBD5 and Notification-Value = 1 with the PLSP-ID(i) it associated to this inter-domain path. For the Backward Recursive method, PCE(i), for i-1 to 1, MUST send backward the PCNtf message to PCE(i-1) up to PCE(1) by replacing the received PLSP-ID(i+1) by the PLSP-ID(i) it reported previously.

When a PCE(i) would withdraw from an inter-domain path, it MUST send a PCNtf message backward to PCE(1) for the Backward Recursive method or to parent PCE for the Hierarchical method with Notification-Type = TBD5 and Notification-Value = 2 with the PLS-ID(i) associated to this inter-domain path. Propagation of the PCNtf message up to PCE(1) in the Backward Recursive method is done as previously.

When a PCE(1) for the Backward Recursive method or parent PCE for the Hierarchical method receives a PCNtf message with Inter-domain Type and Notification-Value set to 1 or 2, it MUST trigger a new path computation for the inter-domain path identified by the PLSP-ID contained in the PCNtf message and MUST update the inter-domain path accordingly.

When an inter-domain path has been updated, following a PCE(1) or a parent PCE initiative or subsequent to an Inter-Domain PCNtf message with Notification-Value = 1 or 2, the PCE(1) for the Backward Recursive method or parent PCE for the Hierarchical method MUST send to PCE(i), for i = 1 to n, a PCNtf message with Notification-Type = TBD5 and Notification-Value = 3 with the PLSP-ID of the inter-domain path. For Backward Recursive method, PCE(i), for i = 2 ... n, MUST replace the PLSP-ID(i) by the PLSP-ID(i+1) they previously associated for this inter-domain path before forwarding this PCNtf message to PCE(i+1). For the hierarchical method, the parent PCE MUST send the PCNtf message with the PLSP-ID(i) reported previously by the PCE(i).

If a PCE received an Inter-Domain PCNtf message with Notification-Value different from the defined value or without the LSP-ID TLV, it MUST reject the message.

5.4.2. Inter-Domain path update

If the PCE(1) or the parent PCE would update the inter-domain path, following the reception of a PCNtf message with Notification-Type = TBD5 and Notification-Value = 1 or 2 or due to a management operation, it MUST respect the following sequence:

* Start a new end-to-end path computation, by means of BRPC for the Backward Recursive method or by contacting Child PCE(i) for the Hierarchical method,

* Once the new path computed, PCE(1) MUST send a PCUpd message to next PCE in the chain and PCE(i) MUST propagate the PCUpd message up to the destination domain PCE(n) for the Backward Recursive method or parent PCE MUST send the PCUpd message to the destination domain Child PCE(n) for the Hierarchical method.

* Once destination domain reached, each PCE(i) for i = n to 1 MUST execute the following steps depending if the border nodes are modified or not:

* - BN-en(i), BN-ex(i) and BN-en(i+1) are not modified: PCUpd message must be used and PCE(i) MUST execute the following steps:

  - o PCE(i) MUST send the PCUpd message to the PCC BN-en(i) node with the new path characteristics

    o The PCC BN-en(i) process the PCUpd message like for PCInitiate message (see section 3.1 or section 4.1) and send back a Stitching Label within the PCRpt message to its PCE(i). Note that the PCC BN-en(i) MAY return a different Stitching Label value.

    o Once the PCE(i) received the PCRpt message from the PCC BN-en(i) with the Stitching Label, it MUST send back a PCRpt message to the previous PCE(i-1) for the Backward Recursive method or to parent PCE for the Hierarchical method to acknowledge that the PCUpd message has been processed with the PLSP-ID it allocated previously for its part of the inter-domain path.

- BN-en(i), BN-ex(i) and/or BN-en(i+1) are modified: PCUpd message could not be used as source and/or destination of the local part of the inter-domain path has changed. PCInitiate message must be used and PCE(i) MUST execute the following steps:

- o PCE(i) MUST send a PCInitiate message to the updated PCC BN-en(i) node to create the new path between the updated BN-en(i) and BN-en(i+1) nodes through updated BN-ex(i) with the requested Binding SID Object as described previously in section 3.1 or 4.1,

  o PCC BN-en(i) process the PCInitiate message as described in section 3.1 or section 4.1 and send back a Stitching Label within the PCRpt message to its PCE(i),

  o Once PCE(i) received the PCRpt message with the Stitching Label, it MUST first associate the new PLSP-ID with the PLSP-ID it allocated previously and reported to PCE(i-1) and keep the old PLSP-ID of the old PCC BN-en(i) for further reference. Then, it MUST send back a PCRpt message to the previous PCE(i-1) for the Backward Recursive method or to parent PCE for the Hierarchical method to acknowledge that the update has been processed with the PLSP-ID it allocated previously for its part of the inter-domain path,

  o PCE(i+1) MUST request a new Stitching Label to its PCC BN-en(i+1) if the BN-ex(i) has been modified in order to associated the correct Stitching Label with the new Link ID for the incoming traffic even if border nodes are not modified in its domain. Then it MUST send a PCRpt message with the new Stitching Value to PCE(i) in order for the latter to steer the traffic to the new path.

* For the Backward Recursive method, when PCE(1) received the PCRpt message from its PCC BN-en(1) in response to the PCUpd message, it MUST send a PCNtf message with the Notification-Type = TBD5 and Notification-Value = 2 to advertise the PCE(i), for i = 2 to n, that the inter-domain path has been updated:

* - PCE(i), for i = 1 to n, MUST remove old local path that is no more part of the inter-domain path following the update in the case where BN-en(i), BN-ex(i) and/or BN-en(i+1) have been modified by sending a PCInitiate message with R flag set to 1 to the PCC BN-en(i). PCE(i) could use the old PLSP-ID it keeps during the update phase to help the removal of this old path,

- PCE(i) MUST update the Inter-Domain Association group with the new PLSP-ID,

- Then, it MUST forward the PCNtf message received to PCE(i+1) for the Backward Recursive method up to PCE(n).

* For the Hierarchical method, when parent PCE received the PCRpt message from its Child PCE(1) in response to its PCUpd message, it MUST send a PCNtf message with the Notification-Type = TBD5 and Notification-Value = 2 to advertise the Child PCE(i), for i = 1..n, that the inter-domain path has been updated:

* - Child PCE(i), for i = 1 to n, MUST remove old local path that is no more part of the inter-domain path following the update in the case where BN-en(i), BN-ex(i) and/or BN-en(i+1) have been modified by sending a PCInitiate message with R flag set to 1 to the PCC BN-en(i). PCE(i) could use the old PLSP-ID it keeps during the update phase to help the removal of this old path,

- PCE(i) MUST update the Inter-Domain Association group with the new PLSP-ID.

The Inter-Domain Association Object MUST be present in the PCUpd message in the Backward Recursive method and MAY be present in the Hierarchical method.

5.5. Modification of Local Paths

The modification of local paths, i.e. between BN-en(i) and BN-ex(i), is left to the discretion of PCE(i). More precisely, if the PCE(i) wishes to modify the local part of the inter-domain path, it MUST send a standard PCUpd message and wait to receive the corresponding PCRpt message. 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 message indicates that the Stitching Label SL(i) has changed.

However, in the case of modification of the local path of inter-domain paths, the PCE(i) MUST respect following policy rules:

* PCE(i) MUST NOT modify the BN-ex(i-1), BN-en(i), BN-ex(i) and BN-en(i+1) as the inter-domain path is only delegated to the PCE(i-1) and is not initiated by the PCE(i),

* PCE(i) MUST NOT degrade the local path regarding the constraints of the inter-domain path e.g. PCE(i) could not compute a path with larger delay metric or less Bandwidth as initial request,

* The PCE(i) MUST generate a PCRpt message to the PCE(i-1) if and only if the modification of the local part of the inter-domain path affects the domain(i-1), i.e. if the Stitching Label is modified or if the modification failed and the local path goes down

All other modifications that imply to not respect the previous policy rules, MUST follow the modification of inter-domain path as describe in previous section.

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: tear-down operation starts from PCE(n) backward to PCE(1) through all PCE(i).

* 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: tear-down operation starts from PCE(n) backward to PCE(1) through all PCE(i).

Each domain PCE(i) is responsible to tear down its part of the path and the PCC MUST release both the Stitching label SL(i) 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(i). The Association Group MUST also be removed by the PCC and PCE(i). When PCE(i) received back the PCRpt message from the PCC and after removing all context, it sends back a PCRpt message to PCE(i-1) to indicate that all local paths and contexts have been correctly removed. This will trigger tear-down operation on PCE(i-1) and so on.

5.7. Reporting of Inter-domain path status

In stateful mode, PCC must report any new modification of the state of LSP to their PCE by sending an appropriate PCRpt message.

In the case of inter-domain path, when a PCE(i) receives a new PCRpt message from a PCC BN-en(i), it MUST send is back to the PCE(i-1) with the PLSP-ID it reported previously in the Backward Recursive method and to the parent PCE in the Hierarchical method if and only if it impacts the state of the inter-domain path.

In case a failure appears 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 an inter-domain path update as described in section 5.4.

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 IP/MPLS networks are supported by this specification.

6.1. 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 an LSP containing an empty TE-PATH-BINDING TLV with the I flag set to 1 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, Segment Routing or SRv6. Following sections is described how the mapping between technology take place:

6.1.1. RSVP-TE and SR-MPLS

The PST value 0 or 1, which indicates if the local segment path is RSVP-TE or SR-MPLS, 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 used here the same technology i.e. MPLS Label and 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. Finally, PCE(i) MUST complete accordingly ERO(i) with the identifier of Link(i+1): IP address of link between BN-ex(i) and BN-en(i+1) for RSVP-TE or EPE label of link between BN-ex(i) and BN-en(i+1) for Segment Routing.

6.1.2. SRv6 and MPLS

Binding Segment (BSID) is bound to SR Policy [RFC8402]. Thus, an SR-MPLS label can be bound to an SRv6 Policy and an SRv6 SID can be bound to an SR-MPLS Policy which allow such interworking.SRv6 and MPLS interworking draft [I-D.ietf-spring-srv6-mpls-interworking] defines how SRv6 and MPLS dataplane could coexist and how control plane should be used to coordinate both technology. In particular, section 7.1.1 defines precisely how SR-PCE could be used. It makes the distinction between 6oM (SRv6 to MPLS) and Mo6 (MPLS to SRv6) interworking. However, the interworking use case corresponds to core domain that interconnect same technology i.e. SRv6 crossing an MPLS domain or MPLS crossing an SRv6 domain while, in this document, each domains are independant and we address how interworking is possible between two adjacent domains independently of service or other

domain.

6.1.2.1. 6toM

Packets are processed as described in section 7.1.1.1 (6oM) of Srv6 and MPLS interworking draft [I-D.ietf-spring-srv6-mpls-interworking]. However, two scenarios should be distinguished depending which domain provides the gateway for the mapping between SR Policy:

* Gateway is router BN-ex(i): In this case, PCE(i) creates an SRv6 Policy on node BN-en(i) with an SRH bounded to an BN-en(i+1) Endpoint End.BM (instead of End.B6) associated with the MPLS Stitching Label SL(i+1) as argument. When BN-ex(i) receives a packet with such SRv6 Policy, it removes the SRH header and push the Stiching Label SL(i+1) on front of the packet and send it to the BN-en(i+1) router. Then, BN-en(i+1) processes packet as described previously.

* Gateway is router BN-en(i+1): In this case, PCE(i) creates an SRv6 Policy on node BN-en(i) with an SRH bounded to an BN-en(i+1) Endpoint End.X to indicates the inter-domain link followed by an Endpoint End.BM associated with the MPLS Stitching Label SL(i+1) as argument. The router BN-ex(i) selects next SRH header i.e. End.X and sends packet to BN-en(i+1). In turn, BN-en(i+1) removes the SRH and then look to its LFIB with the Stitching Label SL(i+1) as entry to determine which SR-MPLS Policy it should apply: i.e. the SL(i+1) point to the stack of labels in the case of SR-MPLS or the label of the RSVP-TE tunnel.

Selection of one of the two scenarios will depend of the presence of S and 6 flag in PCEP Capability as defined in section 5.1. However, the first scenario don’t need to implement dual technology in the Border Nodes but loose the possibility to select the inter-domain link while the second scenario allows a fine selection of inter-domain link but impose a dual technology support for the router BN-en(i+1).

6.1.2.2. Mto6

Packets are processed as described in section 7.1.1.2 (Mo6) of Srv6 and MPLS interworking draft [I-D.ietf-spring-srv6-mpls-interworking]. Again, two scenarios take place depending which domain provides the mapping between SR Policy:

* Gateway is router BN-ex(i): In this case, PCE(i) creates an SR-MPLS Policy on BN-en(i) bounded by a local Binding Label. It also creates an SRv6 Policy on BN-ex(i) associated to this Binding Label. When BN-ex(i) receives the packets, it first pops the last

MPLS label i.e. the local Binding Label and encapsulates the packets with and SRH header which corresponds to the BN-en(i+1) SRv6 SID with the End.B6.encaps behavior and an argument that corresponds to the Stitching Label SL(i+1). Then, it forward the packets to BN-en(i+1) as usual and process continue as described previously.

* Gateway is router BN-en(i+1): In this case, PCE(i) creates an SR-MPLS Policy on BN-en(i) bounded by Adjacency SID for the inter-domain link and the Stitching Label SL(i+1). When BN-ex(i) receives the packets encapsulated with the Adjacency SID, it pops this label and send the packet with the last label in the stack i.e. the Stitching Label SL(i+1) to BN-en(i+1). When BN-en(i+1) receives the packet, it must pops this label and replaces it by the SRH header that correspond to the SRv6 Policy created previously by PCE(i+1). Then it forward the packets according to the SRv6 Policy as usual.

Scenarios are govern by the presence of S and 6 flags in the PCE Capability defined in section 5.1. Compared to 6toM use case, the second scenario is preferred as it is simpler for an SRv6 aware router to use an MPLS label to steer traffic into an SRv6 segment path. First scenario loose the possibility of fine selection of inter-domain link and it impose to PCE(i) to setup both and SR-MPLS and SRv6 Policy in its domain.

6.2. 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 exceed the MSD. The SL could be used to split a too long explicit path regarding the MSD constraints. In this scenario, there are also no communications between PCEs and new TLVs are only used between PCE and PCCs.

6.3. Nested traffic

When a domain(i) would group into the same local path all traffic that entering into the domain through the same border node BN-en(i) and exit by the same border node BN-ex(i), it could be useful to identify the different inter-domain paths within this local path. Indeed, traffic entering in this nested local path could goes to different domains or different destination of the same domain. Thus, it is mandatory to keep them perfectly identifiable through a dedicated Stitching Label. In this case, PCE(i) proceeds as if it nested internal traffic. Nested tunnel MUST be created in top of existing inter-domain local path. Subsequent inter-domain local path will follow this nested local path. As a consequence, PCE(i) MUST NOT request a second Stitching Label(i) for an existing inter-domain local path.

6.4. VPN

For VPN use case, inter-domain paths link PEs that spawn in different domains. Such connectivity could be considered as Seamless MPLS. In order for the source PE to route L3/L2 VPN packets to the destination, it must be aware that the next-hop BGP i.e. the destination PE, is joinable through the appropriate inter-domain path. For that purpose, the corresponding prefix i.e. the L3/L2 VPN one, must be exchange with ’next hop unchanged’ command to keep unmodified the IP address of the PE which advertised this VPN prefix. And like for inter-domain option C, Route Reflectors could be used to exchange and advertise between domains the IP addresses of BGP next-hop i.e. in general the loopback IP address of the PE routers.

6.5. Intent-Based Networking

Intent-Based Networking as per [RFC9315] defines goals and outcomes that an implementation must respect. The present memo addresses most of the concepts of Intent-Based Networking and in particular:

* From a service point of view, the inter-domain path computation is abstracted though by the use of PKS and enforcement technology i.e. RSVP-TE, SR-TE, over-provisioning ... is done independently by each domain. The user is therefore unaware of the details of the creation, modification, deletion, and management of the inter-domain path,

* AS path computation performed during the initial step allows to include or exclude some domains allowing geography or geo-political routing,

* Intra-domain path computation done by domain(i) allows to include or exclude some Border Nodes, internal nodes, internal links ... based on any policy deployed by domain(i) and independently of the other domains,

* Traffic usage is controlled by each domain independently while end-to-end constraints are respected during inter-domain path computation and enforcement.

In addition, the Backward Recursive method avoids the usage of a centralize controller and completely distributes the service instantiation across the involved domains.

6.6. QoS management

When performing the path computation, the initial request to the PCE(1) could contains the CLASSTYPE (CT) Object in the list of constrained parameters. This Class-Type MUST be propagated during the inter-domain path computation in the Backward Recursive method, respectively to the Parent PCE in the Hierarchical method and take into account for Bandwidth Reservation. As per [RFC5455], each PCE(i) MUST store the Class-Type in order to complete the path computation, in particular in the Backward Recursive method. In addition, the Class-Type MAY be used by domain(i) to set the EXP bits in the MPLS label in order for router to process the packets accordingly to DiffServ-TE configured in the domain(i). To maintain the independence, domain(i) is free to implement or not DiffServ-TE queuing, and if DiffServe-TE queuing is implemented, to configure the different queue parameters with its own traffic engineering rules. Thus, the proposed mechanism just allows to propagate the EXP bits without modification. If domain(i) is not able or would not accept traffic in the specified Class-Type, during the inter-domain path computation it MUST reply with a NO_PATH Object to domain(i-1) for the Backward Recursive method, respectively to the Parent PCE in the Hierarchical method.

In case of Bandwidth Reservation i.e. initial request constraints a Bandwidth Metric Object, domain (i) MAY configure the BN-en(i) of the inter-domain path to verify and enforce traffic corresponding to the requested bandwidth with the mechanism domain(i) seems appropriate. In other words, domain(i) MAY apply its own traffic engineering and policy rules at the entry of its domain independently from the other domains. The detailed configuration of the QoS for the BN-en(i) router is outside the scope of this draft.

7. IANA Considerations

7.1. TE-PATH-BINDING flag

Binding Label / Segment Identifier (BSID) [RFC9604] defines the TE-PATH-BINDING TLV Flag field. IANA is requested to allocate new flag in the PCEP TE-PATH-BINDING TLV Flags field registry, as follows:

Bit	Description	Reference
1	I (Inter-domain Binding Label/Segment Identifier)	This Document
2	S (Strictly steer traffic)	This Document

Table 1

7.2. 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 6 Mandatory Object Missing Error and new error-value of Error-Type 26 Association Error:

Error-Type	Error-Value	Description
6	TBD1	LSP TE-PATH-BINDING missing TLV
26	TBD2	Error in association of Inter-domain LSPs

Table 2

7.3. PCEP TLV Type Indicators

IANA is requested to allocate a new TLV Type Indicator for the "Inter-Domain PCE Capability" within the "PCEP TLV Type Indicators" sub-registry of the "Path Computation Element Protocol (PCEP) Numbers" registry:

Value	Description	Reference
TBD3	INTER-DOMAIN-PCE-CAPABILITY	This Document

Table 3

7.4. Inter-Domain PCE Capability

IANA is requested to allocate a new sub-registry, named "INTER-DOMAIN-PCE-CAPABILITY TLV Flag Field", within the "Path Computation Element Protocol (PCEP) Numbers" registry, to manage the Flag field in the INTER-DOMAIN-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	R: RECURSIVE-PATH-COMPUTATION-CAPABILITY	This Document
30	S: INTER-DOMAIN-STITCHING-LABEL-CAPABILITY	This Document
29	6: INTER-DOMAIN-STITCHING-SRV6-CAPABILITY	This Document

Table 4

7.5. H-PCE Capability

H-PCE [RFC8685] defines the PCEP H-PCE-CAPABILITY TLV Flags field. IANA is requested to allocate new flag in the PCEP H-PCE-CAPABILITY TLV Flag field registry, as follows:

Bit	Description	Reference
30	S: INTER-DOMAIN-STITCHING-LABEL-CAPABILITY	This Document
29	6: INTER-DOMAIN-STITCHING-SRV6-CAPABILITY	This Document

Table 5

7.6. 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

Table 6

7.7. PCEP Notify Type and Values

IANA is requested to allocate code-points in the PCEP-NOTIFICATION Object registry for a new notification type named "Inter-domain Path" and new notification values within this Notification-Type as follow:

Notification-Type	Notification-Value	Description
TBD5	1	This Inter-domain path needs update
	2	This domain is no more usable for this Inter-domain path
	3	This Inter-domain path has been updated

Table 7

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. Acknowledgments

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>.

[RFC8685] Zhang, F., Zhao, Q., Gonzalez de Dios, O., Casellas, R., and D. King, "Path Computation Element Communication Protocol (PCEP) Extensions for the Hierarchical Path Computation Element (H-PCE) Architecture", RFC 8685, DOI 10.17487/RFC8685, December 2019, <https://www.rfc-editor.org/info/rfc8685>.

[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>.

[RFC9604] Sivabalan, S., Filsfils, C., Tantsura, J., Previdi, S., and C. Li, Ed., "Carrying Binding Label/SID in PCE-Based Networks", RFC 9604, DOI 10.17487/RFC9604, August 2024, <https://www.rfc-editor.org/info/rfc9604>.

11.2. Informative References

[I-D.ietf-spring-srv6-mpls-interworking]
Agrawal, S., Filsfils, C., Voyer, D., Dawra, G., Li, Z., and S. Hegde, "SRv6 and MPLS interworking", Work in Progress, Internet-Draft, draft-ietf-spring-srv6-mpls-interworking-00, 17 October 2024, <https://datatracker.ietf.org/doc/html/draft-ietf-spring-srv6-mpls-interworking-00>.

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

[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>.

[RFC5455] Sivabalan, S., Ed., Parker, J., Boutros, S., and K. Kumaki, "Diffserv-Aware Class-Type Object for the Path Computation Element Communication Protocol", RFC 5455, DOI 10.17487/RFC5455, March 2009, <https://www.rfc-editor.org/info/rfc5455>.

[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>.

[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>.

[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>.

[RFC8986] Filsfils, C., Ed., Camarillo, P., Ed., Leddy, J., Voyer, D., Matsushima, S., and Z. Li, "Segment Routing over IPv6 (SRv6) Network Programming", RFC 8986, DOI 10.17487/RFC8986, February 2021, <https://www.rfc-editor.org/info/rfc8986>.

[RFC9086] Previdi, S., Talaulikar, K., Ed., Filsfils, C., Patel, K., Ray, S., and J. Dong, "Border Gateway Protocol - Link State (BGP-LS) Extensions for Segment Routing BGP Egress Peer Engineering", RFC 9086, DOI 10.17487/RFC9086, August 2021, <https://www.rfc-editor.org/info/rfc9086>.

[RFC9315] Clemm, A., Ciavaglia, L., Granville, L. Z., and J. Tantsura, "Intent-Based Networking - Concepts and Definitions", RFC 9315, DOI 10.17487/RFC9315, October 2022, <https://www.rfc-editor.org/info/rfc9315>.

[RFC9514] Dawra, G., Filsfils, C., Talaulikar, K., Ed., Chen, M., Bernier, D., and B. Decraene, "Border Gateway Protocol - Link State (BGP-LS) Extensions for Segment Routing over IPv6 (SRv6)", RFC 9514, DOI 10.17487/RFC9514, December 2023, <https://www.rfc-editor.org/info/rfc9514>.

Authors' Addresses

Olivier Dugeon
Orange Innovation
2, Avenue Pierre Marzin
22307 Lannion
France
Email: olivier.dugeon@orange.com

Julien Meuric
Orange Innovation
2, Avenue Pierre Marzin
22307 Lannion
France
Email: julien.meuric@orange.com

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

Daniele Ceccarelli
Cisco
Via S. M. Molgora, 48/C,
20871 Vimercate
Italy
Email: daniele.ietf@gmail.com