Internet Engineering Task Force (IETF)
Request for Comments: 8233
Category: Standards Track
ISSN: 2070-1721

D. Dhody
Q. Wu
Huawei
V. Manral
Nano Sec Co
Z. Ali
Cisco Systems
K. Kumaki
KDDI Corporation

September 2017

Extensions to the Path Computation Element Communication Protocol (PCEP) to Compute Service-Aware Label Switched Paths (LSPs)

Abstract

In certain networks, such as, but not limited to, financial information networks (e.g., stock market data providers), network performance criteria (e.g., latency) are becoming as critical to data path selection as other metrics and constraints. These metrics are associated with the Service Level Agreement (SLA) between customers and service providers. The link bandwidth utilization (the total bandwidth of a link in actual use for the forwarding) is another important factor to consider during path computation.

IGP Traffic Engineering (TE) Metric Extensions describe mechanisms with which network performance information is distributed via OSPF and IS-IS, respectively. The Path Computation Element Communication Protocol (PCEP) provides mechanisms for Path Computation Elements (PCEs) to perform path computations in response to Path Computation Client (PCC) requests. This document describes the extension to PCEP to carry latency, delay variation, packet loss, and link bandwidth utilization as constraints for end-to-end path computation.

Status of This Memo

Table of Contents

1. Introduction 3
   1.1. Requirements Language 4
2. Terminology 4
3. PCEP Extensions 5
   3.1. Extensions to METRIC Object 5
        3.1.1. Path Delay Metric 6
               3.1.1.1. Path Delay Metric Value 7
        3.1.2. Path Delay Variation Metric 7
               3.1.2.1. Path Delay Variation Metric Value 8
        3.1.3. Path Loss Metric 8
               3.1.3.1. Path Loss Metric Value 9
        3.1.4. Non-Understanding / Non-Support of
               Service-Aware Path Computation 9
        3.1.5. Mode of Operation 10
               3.1.5.1. Examples 11
        3.1.6. Point-to-Multipoint (P2MP) 11
               3.1.6.1. P2MP Path Delay Metric 11
               3.1.6.2. P2MP Path Delay Variation Metric 12
               3.1.6.3. P2MP Path Loss Metric 12
   3.2. Bandwidth Utilization 12
        3.2.1. Link Bandwidth Utilization (LBU) 12
        3.2.2. Link Reserved Bandwidth Utilization (LRBU) 13
        3.2.3. Bandwidth Utilization (BU) Object 13
               3.2.3.1. Elements of Procedure 14
   3.3. Objective Functions 15
4. Stateful PCE and PCE Initiated LSPs 16
5. PCEP Message Extension 17
   5.1. The PCReq Message 17
   5.2. The PCRep Message 18
   5.3. The PCRpt Message 19
6. Other Considerations 20

6.1. Inter-domain Path Computation 20
6.1.1. Inter-AS Links 20
6.1.2. Inter-Layer Path Computation 20
6.2. Reoptimizing Paths 21
7. IANA Considerations 21
7.1. METRIC Types 21
7.2. New PCEP Object 22
7.3. BU Object 22
7.4. OF Codes 22
7.5. New Error-Values 23
8. Security Considerations 23
9. Manageability Considerations 24
9.1. Control of Function and Policy 24
9.2. Information and Data Models 24
9.3. Liveness Detection and Monitoring 24
9.4. Verify Correct Operations 24
9.5. Requirements on Other Protocols 24
9.6. Impact on Network Operations 24
10. References 25
10.1. Normative References 25
10.2. Informative References 26
Appendix A. PCEP Requirements 28
Acknowledgments 29
Contributors 30
Authors' Addresses 31

1. Introduction

Real-time network performance information is becoming critical in the path computation in some networks. Mechanisms to measure latency, delay variation, and packet loss in an MPLS network are described in [RFC6374]. It is important that latency, delay variation, and packet loss are considered during the path selection process, even before the Label Switched Path (LSP) is set up.

Link bandwidth utilization based on real-time traffic along the path is also becoming critical during path computation in some networks. Thus, it is important that the link bandwidth utilization is factored in during the path computation.

The Traffic Engineering Database (TED) is populated with network performance information like link latency, delay variation, packet loss, as well as parameters related to bandwidth (residual bandwidth, available bandwidth, and utilized bandwidth) via TE Metric Extensions in OSPF [RFC7471] or IS-IS [RFC7810] or via a management system. [RFC7823] describes how a Path Computation Element (PCE) [RFC4655] can use that information for path selection for explicitly routed LSPs.

A Path Computation Client (PCC) can request a PCE to provide a path meeting end-to-end network performance criteria. This document extends the Path Computation Element Communication Protocol (PCEP) [RFC5440] to handle network performance constraints that include any combination of latency, delay variation, packet loss, and bandwidth utilization constraints.

[RFC7471] and [RFC7810] describe various considerations regarding:

- o Announcement thresholds and filters
- o Announcement suppression
- o Announcement periodicity and network stability

The first two provide configurable mechanisms to bound the number of re-advertisements in IGP. The third provides a way to throttle announcements. Section 1.2 of [RFC7823] also describes the oscillation and stability considerations while advertising and considering service-aware information.

1.1. Requirements Language

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

2. Terminology

The following terminology is used in this document.

IGP: Interior Gateway Protocol; either of the two routing protocols, Open Shortest Path First (OSPF) or Intermediate System to Intermediate System (IS-IS).

IS-IS: Intermediate System to Intermediate System

LBU: Link Bandwidth Utilization (see Section 3.2.1)

LRBU: Link Reserved Bandwidth Utilization (see Section 3.2.2)

MPLP: Minimum Packet Loss Path (see Section 3.3)

MRUP: Maximum Reserved Under-Utilized Path (see Section 3.3)

MUP: Maximum Under-Utilized Path (see Section 3.3)

OF: Objective Function; a set of one or more optimization criteria used for the computation of a single path (e.g., path cost minimization) or for the synchronized computation of a set of paths (e.g., aggregate bandwidth consumption minimization, etc.). (See [RFC5541].)

OSPF: Open Shortest Path First

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

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

RSVP: Resource Reservation Protocol

TE: Traffic Engineering

TED: Traffic Engineering Database

3. PCEP Extensions

This section defines PCEP extensions (see [RFC5440]) for requirements outlined in Appendix A. The proposed solution is used to support network performance and service-aware path computation.

3.1. Extensions to METRIC Object

The METRIC object is defined in Section 7.8 of [RFC5440], comprising metric-value and metric-type (T field), and a flags field, comprising a number of bit flags (B bit and P bit). This document defines the following types for the METRIC object.

o T=12: Path Delay metric (Section 3.1.1)

o T=13: Path Delay Variation metric (Section 3.1.2)

o T=14: Path Loss metric (Section 3.1.3)

o T=15: P2MP Path Delay metric (Section 3.1.6.1)

o T=16: P2MP Path Delay Variation metric (Section 3.1.6.2)

o T=17: P2MP Path Loss metric (Section 3.1.6.3)

The following terminology is used and expanded along the way.

- o A network comprises of a set of N links {Li, (i=1...N)}.

- o A path P of a point-to-point (P2P) LSP is a list of K links {Lpi,(i=1...K)}.

3.1.1. Path Delay Metric

The Link Delay metric is defined in [RFC7471] and [RFC7810] as "Unidirectional Link Delay". The Path Delay metric type of the METRIC object in PCEP represents the sum of the Link Delay metric of all links along a P2P path. Specifically, extending on the above-mentioned terminology:

- o A Link Delay metric of link L is denoted D(L).

- o A Path Delay metric for the P2P path P = Sum {D(Lpi), (i=1...K)}.

This is as per the sum of means composition function (Section 4.2.5 of [RFC6049]). Section 1.2 of [RFC7823] describes oscillation and stability considerations, and Section 2.1 of [RFC7823] describes the calculation of the end-to-end Path Delay metric. Further, Section 4.2.9 of [RFC6049] states when this composition function may fail.

Metric Type T=12: Path Delay metric

A PCC MAY use the Path Delay metric in a Path Computation Request (PCReq) message to request a path meeting the end-to-end latency requirement. In this case, the B bit MUST be set to suggest a bound (a maximum) for the Path Delay metric that must not be exceeded for the PCC to consider the computed path as acceptable. The Path Delay metric must be less than or equal to the value specified in the metric-value field.

A PCC can also use this metric to ask PCE to optimize the path delay during path computation. In this case, the B bit MUST be cleared.

A PCE MAY use the Path Delay metric in a Path Computation Reply (PCRep) message along with a NO-PATH object in the case where the PCE cannot compute a path meeting this constraint. A PCE can also use this metric to send the computed Path Delay metric to the PCC.

3.1.1.1. Path Delay Metric Value

[RFC7471] and [RFC7810] define "Unidirectional Link Delay Sub-TLV" to advertise the link delay in microseconds in a 24-bit field. [RFC5440] defines the METRIC object with a 32-bit metric value encoded in IEEE floating point format (see [IEEE.754]). Consequently, the encoding for the Path Delay metric value is quantified in units of microseconds and encoded in IEEE floating point format. The conversion from 24-bit integer to 32-bit IEEE floating point could introduce some loss of precision.

3.1.2. Path Delay Variation Metric

The Link Delay Variation metric is defined in [RFC7471] and [RFC7810] as "Unidirectional Delay Variation". The Path Delay Variation metric type of the METRIC object in PCEP encodes the sum of the Link Delay Variation metric of all links along the path. Specifically, extending on the above-mentioned terminology:

o A delay variation of link L is denoted DV(L) (average delay variation for link L).

o A Path Delay Variation metric for the P2P path P = Sum {DV(Lpi), (i=1...K)}.

Section 1.2 of [RFC7823] describes oscillation and stability considerations, and Section 2.1 of [RFC7823] describes the calculation of the end-to-end Path Delay Variation metric. Further, Section 4.2.9 of [RFC6049] states when this composition function may fail.

Note that the IGP advertisement for link attributes includes the average delay variation over a period of time. An implementation, therefore, MAY use the sum of the average delay variation of links along a path to derive the delay variation of the path. An end-to-end bound on delay variation is typically used as constraint in the path computation. An implementation MAY also use some enhanced composition function for computing the delay variation of a path with better accuracy.

Metric Type T=13: Path Delay Variation metric

A PCC MAY use the Path Delay Variation metric in a PCReq message to request a path meeting the path delay variation requirement. In this case, the B bit MUST be set to suggest a bound (a maximum) for the Path Delay Variation metric that must not be exceeded for the PCC to consider the computed path as acceptable. The path delay variation must be less than or equal to the value specified in the metric-value field.

A PCC can also use this metric to ask the PCE to optimize the path delay variation during path computation. In this case, the B flag MUST be cleared.

A PCE MAY use the Path Delay Variation metric in a PCRep message along with a NO-PATH object in the case where the PCE cannot compute a path meeting this constraint. A PCE can also use this metric to send the computed end-to-end Path Delay Variation metric to the PCC.

3.1.2.1. Path Delay Variation Metric Value

[RFC7471] and [RFC7810] define "Unidirectional Delay Variation Sub-TLV" to advertise the link delay variation in microseconds in a 24-bit field. [RFC5440] defines the METRIC object with a 32-bit metric value encoded in IEEE floating point format (see [IEEE.754]). Consequently, the encoding for the Path Delay Variation metric value is quantified in units of microseconds and encoded in IEEE floating point format. The conversion from 24-bit integer to 32-bit IEEE floating point could introduce some loss of precision.

3.1.3. Path Loss Metric

[RFC7471] and [RFC7810] define "Unidirectional Link Loss". The Path Loss (as a packet percentage) metric type of the METRIC object in PCEP encodes a function of the unidirectional loss metrics of all links along a P2P path. The end-to-end packet loss for the path is represented by this metric. Specifically, extending on the above mentioned terminology:

- The percentage link loss of link L is denoted PL(L).

- The fractional link loss of link L is denoted FL(L) = PL(L)/100.

- The percentage Path Loss metric for the P2P path P = (1 - ((1-FL(Lp1)) * (1-FL(Lp2)) * .. * (1-FL(LpK)))) * 100 for a path P with links Lp1 to LpK.

This is as per the composition function described in Section 5.1.5 of [RFC6049].

Metric Type T=14: Path Loss metric

A PCC MAY use the Path Loss metric in a PCReq message to request a path meeting the end-to-end packet loss requirement. In this case, the B bit MUST be set to suggest a bound (a maximum) for the Path Loss metric that must not be exceeded for the PCC to consider the computed path as acceptable. The Path Loss metric must be less than or equal to the value specified in the metric-value field.

A PCC can also use this metric to ask the PCE to optimize the path loss during path computation. In this case, the B flag MUST be cleared.

A PCE MAY use the Path Loss metric in a PCRep message along with a NO-PATH object in the case where the PCE cannot compute a path meeting this constraint. A PCE can also use this metric to send the computed end-to-end Path Loss metric to the PCC.

3.1.3.1. Path Loss Metric Value

[RFC7471] and [RFC7810] define "Unidirectional Link Loss Sub-TLV" to advertise the link loss in percentage in a 24-bit field. [RFC5440] defines the METRIC object with a 32-bit metric value encoded in IEEE floating point format (see [IEEE.754]). Consequently, the encoding for the Path Loss metric value is quantified as a percentage and encoded in IEEE floating point format.

3.1.4. Non-Understanding / Non-Support of Service-Aware Path Computation

If a PCE receives a PCReq message containing a METRIC object with a type defined in this document, and the PCE does not understand or support that metric type, and the P bit is clear in the METRIC object header, then the PCE SHOULD simply ignore the METRIC object as per the processing specified in [RFC5440].

If the PCE does not understand the new METRIC type, and the P bit is set in the METRIC object header, then the PCE MUST send a PCEP Error (PCErr) message containing a PCEP-ERROR Object with Error-Type = 4 (Not supported object) and Error-value = 4 (Unsupported parameter) [RFC5440][RFC5441].

If the PCE understands but does not support the new METRIC type, and the P bit is set in the METRIC object header, then the PCE MUST send a PCErr message containing a PCEP-ERROR Object with Error-Type = 4 (Not supported object) with Error-value = 5 (Unsupported network performance constraint). The path computation request MUST then be canceled.

If the PCE understands the new METRIC type, but the local policy has been configured on the PCE to not allow network performance constraint, and the P bit is set in the METRIC object header, then the PCE MUST send a PCErr message containing a PCEP-ERROR Object with Error-Type = 5 (Policy violation) with Error-value = 8 (Not allowed network performance constraint). The path computation request MUST then be canceled.

3.1.5. Mode of Operation

As explained in [RFC5440], the METRIC object is optional and can be used for several purposes. In a PCReq message, a PCC MAY insert one or more METRIC objects:

- o To indicate the metric that MUST be optimized by the path computation algorithm (path delay, path delay variation, or path loss).

- o To indicate a bound on the METRIC (path delay, path delay variation, or path loss) that MUST NOT be exceeded for the path to be considered as acceptable by the PCC.

In a PCRep message, the PCE MAY insert the METRIC object with an Explicit Route Object (ERO) so as to provide the METRIC (path delay, path delay variation, or path loss) for the computed path. The PCE MAY also insert the METRIC object with a NO-PATH object to indicate that the metric constraint could not be satisfied.

The path computation algorithmic aspects used by the PCE to optimize a path with respect to a specific metric are outside the scope of this document.

All the rules of processing the METRIC object as explained in [RFC5440] are applicable to the new metric types as well.

3.1.5.1. Examples

If a PCC sends a path computation request to a PCE where the metric to optimize is the path delay and the path loss must not exceed the value of M, then two METRIC objects are inserted in the PCReq message:

o First METRIC object with B=0, T=12, C=1, metric-value=0x0000

o Second METRIC object with B=1, T=14, metric-value=M

As per [RFC5440], if a path satisfying the set of constraints can be found by the PCE and there is no policy that prevents the return of the computed metric, then the PCE inserts one METRIC object with B=0, T=12, metric-value= computed path delay. Additionally, the PCE MAY insert a second METRIC object with B=1, T=14, metric-value=computed path loss.

3.1.6. Point-to-Multipoint (P2MP)

This section defines the following types for the METRIC object to be used for the P2MP TE LSPs.

3.1.6.1. P2MP Path Delay Metric

The P2MP Path Delay metric type of the METRIC object in PCEP encodes the Path Delay metric for the destination that observes the worst delay metric among all destinations of the P2MP tree. Specifically, extending on the above-mentioned terminology:

o A P2MP tree T comprises a set of M destinations {Dest_j, (j=1...M)}.

o The P2P Path Delay metric of the path to destination Dest_j is denoted by PDM(Dest_j).

o The P2MP Path Delay metric for the P2MP tree T = Maximum {PDM(Dest_j), (j=1...M)}.

The value for the P2MP Path Delay metric type (T) = 15.

3.1.6.2. P2MP Path Delay Variation Metric

The P2MP Path Delay Variation metric type of the METRIC object in PCEP encodes the Path Delay Variation metric for the destination that observes the worst delay variation metric among all destinations of the P2MP tree. Specifically, extending on the above-mentioned terminology:

- o A P2MP tree T comprises a set of M destinations {Dest_j, (j=1...M)}.
- o The P2P Path Delay Variation metric of the path to the destination Dest_j is denoted by PDVM(Dest_j).
- o The P2MP Path Delay Variation metric for the P2MP tree T = Maximum {PDVM(Dest_j), (j=1...M)}.

The value for the P2MP Path Delay Variation metric type (T) = 16.

3.1.6.3. P2MP Path Loss Metric

The P2MP Path Loss metric type of the METRIC object in PCEP encodes the path packet loss metric for the destination that observes the worst packet loss metric among all destinations of the P2MP tree. Specifically, extending on the above-mentioned terminology:

- o A P2MP tree T comprises of a set of M destinations {Dest_j, (j=1...M)}.
- o The P2P Path Loss metric of the path to destination Dest_j is denoted by PLM(Dest_j).
- o The P2MP Path Loss metric for the P2MP tree T = Maximum {PLM(Dest_j), (j=1...M)}.

The value for the P2MP Path Loss metric type (T) = 17.

3.2. Bandwidth Utilization

3.2.1. Link Bandwidth Utilization (LBU)

The LBU on a link, forwarding adjacency, or bundled link is populated in the TED ("Unidirectional Utilized Bandwidth Sub-TLV" in [RFC7471] and [RFC7810]). For a link or forwarding adjacency, the bandwidth utilization represents the actual utilization of the link (i.e., as measured in the router). For a bundled link, the bandwidth utilization is defined to be the sum of the component link bandwidth utilization. This includes traffic for both RSVP-TE and non-RSVP-TE label switched path packets.

The LBU in percentage is described as the (utilized bandwidth / maximum bandwidth) * 100.

The "maximum bandwidth" is defined in [RFC3630] and [RFC5305] and "utilized bandwidth" in [RFC7471] and [RFC7810].

3.2.2. Link Reserved Bandwidth Utilization (LRBU)

The LRBU on a link, forwarding adjacency, or bundled link can be calculated from the TED. The utilized bandwidth includes traffic for both RSVP-TE and non-RSVP-TE LSPs; the reserved bandwidth utilization considers only the RSVP-TE LSPs.

The reserved bandwidth utilization can be calculated by using the residual bandwidth, available bandwidth, and utilized bandwidth described in [RFC7471] and [RFC7810]. The actual bandwidth by non-RSVP-TE traffic can be calculated by subtracting the available bandwidth from the residual bandwidth ([RFC7471] and [RFC7810]), which is further deducted from utilized bandwidth to get the reserved bandwidth utilization. Thus,

reserved bandwidth utilization = utilized bandwidth - (residual bandwidth - available bandwidth)

The LRBU in percentage is described as the (reserved bandwidth utilization / maximum reservable bandwidth) * 100.

The "maximum reservable bandwidth" is defined in [RFC3630] and [RFC5305]. The "utilized bandwidth", "residual bandwidth", and "available bandwidth" are defined in [RFC7471] and [RFC7810].

3.2.3. Bandwidth Utilization (BU) Object

The BU object is used to indicate the upper limit of the acceptable link bandwidth utilization percentage.

The BU object MAY be carried within the PCReq message and PCRep messages.

BU Object-Class is 35.

BU Object-Type is 1.

The format of the BU object body is as follows:

```
 0                   1                   2                   3
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|              Reserved                         |    Type       |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                     Bandwidth Utilization                     |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

BU Object Body Format

Reserved (24 bits): This field MUST be set to zero on transmission and MUST be ignored on receipt.

Type (8 bits): Represents the bandwidth utilization type. Two values are currently defined.

* Type 1 is LBU (Link Bandwidth Utilization)

* Type 2 is LRBU (Link Residual Bandwidth Utilization)

Bandwidth Utilization (32 bits): Represents the bandwidth utilization quantified as a percentage (as described in Sections 3.2.1 and 3.2.2) and encoded in IEEE floating point format (see [IEEE.754]).

The BU object body has a fixed length of 8 bytes.

3.2.3.1. Elements of Procedure

A PCC that wants the PCE to factor in the bandwidth utilization during path computation includes a BU object in the PCReq message. A PCE that supports this object MUST ensure that no link on the computed path has the LBU or LRBU percentage exceeding the given value.

A PCReq or PCRep message MAY contain multiple BU objects so long as each is for a different bandwidth utilization type. If a message contains more than one BU object with the same bandwidth utilization type, the first MUST be processed by the receiver and subsequent instances MUST be ignored.

If the BU object is unknown/unsupported, the PCE is expected to follow procedures defined in [RFC5440]. That is, if the P bit is set, the PCE sends a PCErr message with error type 3 or 4 (Unknown / Not supported object) and error value 1 or 2 (unknown / unsupported object class / object type), and the related path computation request will be discarded. If the P bit is cleared, the PCE is free to ignore the object.

If the PCE understands but does not support path computation requests using the BU object, and the P bit is set in the BU object header, then the PCE MUST send a PCErr message with a PCEP-ERROR Object Error-Type = 4 (Not supported object) with Error-value = 5 (Unsupported network performance constraint), and the related path computation request MUST be discarded.

If the PCE understands the BU object but the local policy has been configured on the PCE to not allow network performance constraint, and the P bit is set in the BU object header, then the PCE MUST send a PCErr message with a PCEP-ERROR Object Error-Type = 5 (Policy violation) with Error-value = 8 (Not allowed network performance constraint). The path computation request MUST then be canceled.

If path computation is unsuccessful, then a PCE MAY insert a BU object (along with a NO-PATH object) into a PCRep message to indicate the constraints that could not be satisfied.

Usage of the BU object for P2MP LSPs is outside the scope of this document.

3.3. Objective Functions

[RFC5541] defines a mechanism to specify an objective function that is used by a PCE when it computes a path. The new metric types for path delay and path delay variation can continue to use the existing objective function -- Minimum Cost Path (MCP) [RFC5541]. For path loss, the following new OF is defined.

- A network comprises a set of N links {Li, (i=1...N)}.

- A path P is a list of K links {Lpi,(i=1...K)}.

- The percentage link loss of link L is denoted PL(L).

- The fractional link loss of link L is denoted FL(L) = PL(L) / 100.

- The percentage path loss of a path P is denoted PL(P), where PL(P) = (1 - ((1-FL(Lp1)) * (1-FL(Lp2)) * .. * (1-FL(LpK)))) * 100.

Objective Function Code: 9
Name: Minimum Packet Loss Path (MPLP)
Description: Find a path P such that PL(P) is minimized.

Two additional objective functions -- namely, the Maximum Under-Utilized Path (MUP) and the Maximum Reserved Under-Utilized Path (MRUP) are needed to optimize bandwidth utilization. These two new objective function codes are defined below.

These objective functions are formulated using the following additional terminology:

- The bandwidth utilization on link L is denoted u(L).
- The reserved bandwidth utilization on link L is denoted ru(L).
- The maximum bandwidth on link L is denoted M(L).
- The maximum reservable bandwidth on link L is denoted R(L).

The description of the two new objective functions is as follows.

Objective Function Code: 10
Name: Maximum Under-Utilized Path (MUP)
Description: Find a path P such that (Min {(M(Lpi)- u(Lpi)) / M(Lpi), i=1...K }) is maximized.

Objective Function Code: 11
Name: Maximum Reserved Under-Utilized Path (MRUP)
Description: Find a path P such that (Min {(R(Lpi)- ru(Lpi)) / R(Lpi), i=1...K }) is maximized.

These new objective functions are used to optimize paths based on the bandwidth utilization as the optimization criteria.

If the objective functions defined in this document are unknown/ unsupported by a PCE, then the procedure as defined in Section 3.1.1 of [RFC5541] is followed.

4. Stateful PCE and PCE Initiated LSPs

[RFC8231] specifies a set of extensions to PCEP to enable stateful control of MPLS-TE and GMPLS LSPs via PCEP and the maintaining of these LSPs at the stateful PCE. It further distinguishes between an active and a passive stateful PCE. A passive stateful PCE uses LSP state information learned from PCCs to optimize path computations but does not actively update LSP state. In contrast, an active stateful PCE utilizes the LSP delegation mechanism to update LSP parameters in those PCCs that delegated control over their LSPs to the PCE. [PCE-INITIATED] describes the setup, maintenance, and teardown of

PCE-initiated LSPs under the stateful PCE model. The document defines the PCInitiate message that is used by a PCE to request a PCC to set up a new LSP.

The new metric type and objective functions defined in this document can also be used with the stateful PCE extensions. The format of PCEP messages described in [RFC8231] and [PCE-INITIATED] uses <intended-attribute-list> and <attribute-list>, respectively, (where the <intended-attribute-list> is the attribute-list defined in Section 6.5 of [RFC5440] and extended in Section 5.2 of this document) for the purpose of including the service-aware parameters.

The stateful PCE implementation MAY use the extension of PCReq and PCRep messages as defined in Sections 5.1 and 5.2 to enable the use of service-aware parameters during passive stateful operations.

5. PCEP Message Extension

Message formats in this document are expressed using Routing Backus-Naur Form (RBNF) as used in [RFC5440] and defined in [RFC5511].

5.1. The PCReq Message

The extensions to the PCReq message are:

o new metric types using existing METRIC object

o a new optional BU object

o new objective functions using existing OF object [RFC5541]

The format of the PCReq message (with [RFC5541] and [RFC8231] as a base) is updated as follows:

```
   <PCReq Message> ::= <Common Header>
                       [<svec-list>]
                       <request-list>
   where:
        <svec-list> ::= <SVEC>
                        [<OF>]
                        [<metric-list>]
                        [<svec-list>]

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

        <request> ::= <RP>
                      <END-POINTS>
                      [<LSP>]
                      [<LSPA>]
                      [<BANDWIDTH>]
                      [<bu-list>]
                      [<metric-list>]
                      [<OF>]
                      [<RRO>[<BANDWIDTH>]]
                      [<IRO>]
                      [<LOAD-BALANCING>]

   and where:
        <bu-list>::=<BU>[<bu-list>]
        <metric-list> ::= <METRIC>[<metric-list>]
```

5.2. The PCRep Message

The extensions to the PCRep message are:

- new metric types using existing METRIC object
- a new optional BU object (during unsuccessful path computation, to indicate the bandwidth utilization as a reason for failure)
- new objective functions using existing OF object [RFC5541]

The format of the PCRep message (with [RFC5541] and [RFC8231] as a base) is updated as follows:

```
   <PCRep Message> ::= <Common Header>
                       [<svec-list>]
                       <response-list>

   where:

         <svec-list> ::= <SVEC>
                         [<OF>]
                         [<metric-list>]
                         [<svec-list>]

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

        <response> ::= <RP>
                       [<LSP>]
                       [<NO-PATH>]
                       [<attribute-list>]
                       [<path-list>]

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

        <path> ::= <ERO>
                   <attribute-list>

   and where:

        <attribute-list> ::= [<OF>]
                             [<LSPA>]
                             [<BANDWIDTH>]
                             [<bu-list>]
                             [<metric-list>]
                             [<IRO>]

        <bu-list>::=<BU>[<bu-list>]
        <metric-list> ::= <METRIC> [<metric-list>]
```

5.3. The PCRpt Message

A Path Computation LSP State Report message (also referred to as PCRpt message) is a PCEP message sent by a PCC to a PCE to report the current state or delegate control of an LSP. The BU object in a PCRpt message specifies the upper limit set at the PCC at the time of LSP delegation to an active stateful PCE.

The format of the PCRpt message is described in [RFC8231], which uses the <intended-attribute-list>, which is the attribute-list defined in Section 6.5 of [RFC5440] and extended by PCEP extensions.

The PCRpt message can use the updated <attribute-list> (as extended in Section 5.2) for the purpose of including the BU object.

6. Other Considerations

6.1. Inter-domain Path Computation

[RFC5441] describes the Backward Recursive PCE-Based Computation (BRPC) procedure to compute an end-to-end optimized inter-domain path by cooperating PCEs. The new metric types defined in this document can be applied to end-to-end path computation, in a similar manner to the existing IGP or TE metrics. The new BU object defined in this document can be applied to end-to-end path computation, in a similar manner to a METRIC object with its B bit set to 1.

All domains should have the same understanding of the METRIC (path delay variation, etc.) and the BU object for end-to-end inter-domain path computation to make sense. Otherwise, some form of metric normalization as described in [RFC5441] MUST be applied.

6.1.1. Inter-AS Links

The IGP in each neighbor domain can advertise its inter-domain TE link capabilities. This has been described in [RFC5316] (IS-IS) and [RFC5392] (OSPF). The network performance link properties are described in [RFC7471] and [RFC7810]. The same properties must be advertised using the mechanism described in [RFC5392] (OSPF) and [RFC5316] (IS-IS).

6.1.2. Inter-Layer Path Computation

[RFC5623] provides a framework for PCE-based inter-layer MPLS and GMPLS traffic engineering. Lower-layer LSPs that are advertised as TE links into the higher-layer network form a Virtual Network Topology (VNT). The advertisement into the higher-layer network should include network performance link properties based on the end-to-end metric of the lower-layer LSP. Note that the new metrics defined in this document are applied to end-to-end path computation, even though the path may cross multiple layers.

6.2. Reoptimizing Paths

[RFC6374] defines the measurement of loss, delay, and related metrics over LSPs. A PCC can utilize these measurement techniques. In case it detects a degradation of network performance parameters relative to the value of the constraint it gave when the path was set up, or relative to an implementation-specific threshold, it MAY ask the PCE to reoptimize the path by sending a PCReq with the R bit set in the RP object, as per [RFC5440].

A PCC may also detect the degradation of an LSP without making any direct measurements, by monitoring the TED (as populated by the IGP) for changes in the network performance parameters of the links that carry its LSPs. The PCC can issue a reoptimization request for any impacted LSPs. For example, a PCC can monitor the link bandwidth utilization along the path by monitoring changes in the bandwidth utilization parameters of one or more links on the path in the TED. If the bandwidth utilization percentage of any of the links in the path changes to a value less than that required when the path was set up, or otherwise less than an implementation-specific threshold, then the PCC can issue a reoptimization request to a PCE.

A stateful PCE can also determine which LSPs should be reoptimized based on network events or triggers from external monitoring systems. For example, when a particular link deteriorates and its loss increases, this can trigger the stateful PCE to automatically determine which LSPs are impacted and should be reoptimized.

7. IANA Considerations

7.1. METRIC Types

IANA maintains the "Path Computation Element Protocol (PCEP) Numbers" registry at <http://www.iana.org/assignments/pcep>. Within this registry, IANA maintains a subregistry for "METRIC Object T Field". Six new metric types are defined in this document for the METRIC object (specified in [RFC5440]).

IANA has made the following allocations:

Value	Description	Reference
12	Path Delay metric	RFC 8233
13	Path Delay Variation metric	RFC 8233
14	Path Loss metric	RFC 8233
15	P2MP Path Delay metric	RFC 8233
16	P2MP Path Delay variation metric	RFC 8233
17	P2MP Path Loss metric	RFC 8233

7.2. New PCEP Object

IANA maintains Object-Types within the "PCEP Objects" registry. IANA has made the following allocation:

Object Class	Object Type	Name	Reference
35	0	Reserved	RFC 8233
	1	BU	RFC 8233

7.3. BU Object

IANA has created a new subregistry, named "BU Object Type Field", within the "Path Computation Element Protocol (PCEP) Numbers" registry to manage the Type field of the BU object. New values are to be assigned by Standards Action [RFC8126]. Each value should be tracked with the following qualities:

- o Type
- o Name
- o Reference

The following values are defined in this document:

Type	Name	Reference
0	Reserved	RFC 8233
1	LBU (Link Bandwidth Utilization)	RFC 8233
2	LRBU (Link Residual Bandwidth Utilization)	RFC 8233

7.4. OF Codes

IANA maintains the "Objective Function" subregistry (described in [RFC5541]) within the "Path Computation Element Protocol (PCEP) Numbers" registry. Three new objective functions have been defined in this document.

IANA has made the following allocations:

Code Point	Name	Reference
9	Minimum Packet Loss Path (MPLP)	RFC 8233
10	Maximum Under-Utilized Path (MUP)	RFC 8233
11	Maximum Reserved Under-Utilized Path (MRUP)	RFC 8233

7.5. New Error-Values

IANA maintains a registry of Error-Types and Error-values for use in PCEP messages. This is maintained as the "PCEP-ERROR Object Error Types and Values" subregistry of the "Path Computation Element Protocol (PCEP) Numbers" registry.

IANA has made the following allocations:

Two new Error-values are defined for the Error-Type "Not supported object" (type 4) and "Policy violation" (type 5).

Error-Type	Meaning and error values	Reference
4	Not supported object	
	Error-value 5: Unsupported network performance constraint	RFC 8233
5	Policy violation	
	Error-value 8: Not allowed network performance constraint	RFC 8233

8. Security Considerations

This document defines new METRIC types, a new BU object, and new OF codes that do not add any new security concerns beyond those discussed in [RFC5440] and [RFC5541] in itself. Some deployments may find the service-aware information like delay and packet loss to be extra sensitive and could be used to influence path computation and setup with adverse effect. Additionally, snooping of PCEP messages with such data or using PCEP messages for network reconnaissance may give an attacker sensitive information about the operations of the network. Thus, such deployment should employ suitable PCEP security mechanisms like TCP Authentication Option (TCP-AO) [RFC5925] or [PCEPS]. The procedure based on Transport Layer Security (TLS) in [PCEPS] is considered a security enhancement and thus is much better suited for the sensitive service-aware information.

9. Manageability Considerations

9.1. Control of Function and Policy

The only configurable item is the support of the new constraints on a PCE, which MAY be controlled by a policy module on an individual basis. If the new constraint is not supported/allowed on a PCE, it MUST send a PCErr message accordingly.

9.2. Information and Data Models

[RFC7420] describes the PCEP MIB. There are no new MIB Objects for this document.

9.3. Liveness Detection and Monitoring

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

9.4. Verify Correct Operations

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

9.5. Requirements on Other Protocols

The PCE requires the TED to be populated with network performance information like link latency, delay variation, packet loss, and utilized bandwidth. This mechanism is described in [RFC7471] and [RFC7810].

9.6. Impact on Network Operations

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

10. References

10.1. Normative References

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

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

[RFC5305] Li, T. and H. Smit, "IS-IS Extensions for Traffic Engineering", RFC 5305, DOI 10.17487/RFC5305, October 2008, <https://www.rfc-editor.org/info/rfc5305>.

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

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

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

[RFC7471] Giacalone, S., Ward, D., Drake, J., Atlas, A., and S. Previdi, "OSPF Traffic Engineering (TE) Metric Extensions", RFC 7471, DOI 10.17487/RFC7471, March 2015, <https://www.rfc-editor.org/info/rfc7471>.

[RFC7810] Previdi, S., Ed., Giacalone, S., Ward, D., Drake, J., and Q. Wu, "IS-IS Traffic Engineering (TE) Metric Extensions", RFC 7810, DOI 10.17487/RFC7810, May 2016, <https://www.rfc-editor.org/info/rfc7810>.

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

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

10.2. Informative References

[IEEE.754] IEEE, "Standard for Binary Floating-Point Arithmetic", IEEE Standard 754-2008, DOI 10.1109/IEEESTD.2008.4610935, August 2008.

[PCE-INITIATED] Crabbe, E., Minei, I., Sivabalan, S., and R. Varga, "PCEP Extensions for PCE-initiated LSP Setup in a Stateful PCE Model", Work in Progress, draft-ietf-pce-pce-initiated-lsp-10, June 2017.

[PCEPS] Lopez, D., Dios, O., Wu, W., and D. Dhody, "Secure Transport for PCEP", Work in Progress, draft-ietf-pce-pceps-16, September 2017.

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

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

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

[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>.

[RFC5623] Oki, E., Takeda, T., Le Roux, JL., and A. Farrel, "Framework for PCE-Based Inter-Layer MPLS and GMPLS Traffic Engineering", RFC 5623, DOI 10.17487/RFC5623, September 2009, <https://www.rfc-editor.org/info/rfc5623>.

[RFC5925] Touch, J., Mankin, A., and R. Bonica, "The TCP Authentication Option", RFC 5925, DOI 10.17487/RFC5925, June 2010, <https://www.rfc-editor.org/info/rfc5925>.

[RFC6049] Morton, A. and E. Stephan, "Spatial Composition of Metrics", RFC 6049, DOI 10.17487/RFC6049, January 2011, <https://www.rfc-editor.org/info/rfc6049>.

[RFC6374] Frost, D. and S. Bryant, "Packet Loss and Delay Measurement for MPLS Networks", RFC 6374, DOI 10.17487/RFC6374, September 2011, <https://www.rfc-editor.org/info/rfc6374>.

[RFC7420] Koushik, A., Stephan, E., Zhao, Q., King, D., and J. Hardwick, "Path Computation Element Communication Protocol (PCEP) Management Information Base (MIB) Module", RFC 7420, DOI 10.17487/RFC7420, December 2014, <https://www.rfc-editor.org/info/rfc7420>.

[RFC7823] Atlas, A., Drake, J., Giacalone, S., and S. Previdi, "Performance-Based Path Selection for Explicitly Routed Label Switched Paths (LSPs) Using TE Metric Extensions", RFC 7823, DOI 10.17487/RFC7823, May 2016, <https://www.rfc-editor.org/info/rfc7823>.

[RFC8126] Cotton, M., Leiba, B., and T. Narten, "Guidelines for Writing an IANA Considerations Section in RFCs", BCP 26, RFC 8126, DOI 10.17487/RFC8126, June 2017, <https://www.rfc-editor.org/info/rfc8126>.

Appendix A. PCEP Requirements

End-to-end service optimization based on latency, delay variation, packet loss, and link bandwidth utilization are key requirements for service providers. The following associated key requirements are identified for PCEP:

1. A PCE supporting this specification MUST have the capability to compute end-to-end paths with latency, delay variation, packet loss, and bandwidth utilization constraints. It MUST also support the combination of network performance constraints (latency, delay variation, loss,...) with existing constraints (cost, hop-limit,...).

2. A PCC MUST be able to specify any network performance constraint in a PCReq message to be applied during the path computation.

3. A PCC MUST be able to request that a PCE optimizes a path using any network performance criteria.

4. A PCE that supports this specification is not required to provide service-aware path computation to any PCC at any time.

   Therefore, it MUST be possible for a PCE to reject a PCReq message with a reason code that indicates service-aware path computation is not supported. Furthermore, a PCE that does not support this specification will either ignore or reject such requests using pre-existing mechanisms; therefore, the requests MUST be identifiable to legacy PCEs, and rejections by legacy PCEs MUST be acceptable within this specification.

5. A PCE SHOULD be able to return end-to-end network performance information of the computed path in a PCRep message.

6. A PCE SHOULD be able to compute multi-domain (e.g., Inter-AS, Inter-Area, or Multi-Layer) service-aware paths.

Such constraints are only meaningful if used consistently: for instance, if the delay of a computed path segment is exchanged between two PCEs residing in different domains, a consistent way of defining the delay must be used.

Acknowledgments

We would like to thank Alia Atlas, John E. Drake, David Ward, Young Lee, Venugopal Reddy, Reeja Paul, Sandeep Kumar Boina, Suresh Babu, Quintin Zhao, Chen Huaimo, Avantika, and Adrian Farrel for their useful comments and suggestions.

Also, the authors gratefully acknowledge reviews and feedback provided by Qin Wu, Alfred Morton, and Paul Aitken during performance directorate review.

Thanks to Jonathan Hardwick for shepherding this document and providing valuable comments. His help in fixing the editorial and grammatical issues is also appreciated.

Thanks to Christian Hopps for the routing directorate review.

Thanks to Jouni Korhonen and Alfred Morton for the operational directorate review.

Thanks to Christian Huitema for the security directorate review.

Thanks to Deborah Brungard for being the responsible AD.

Thanks to Ben Campbell, Joel Jaeggli, Stephen Farrell, Kathleen Moriarty, Spencer Dawkins, Mirja Kuehlewind, Jari Arkko, and Alia Atlas for the IESG reviews.

Contributors

Clarence Filsfils
Cisco Systems
Email: cfilsfil@cisco.com

Siva Sivabalan
Cisco Systems
Email: msiva@cisco.com

George Swallow
Cisco Systems
Email: swallow@cisco.com

Stefano Previdi
Cisco Systems, Inc
Via Del Serafico 200
Rome 00191
Italy
Email: sprevidi@cisco.com

Udayasree Palle
Huawei Technologies
Divyashree Techno Park, Whitefield
Bangalore, Karnataka 560066
India
Email: udayasree.palle@huawei.com

Avantika
Huawei Technologies
Divyashree Techno Park, Whitefield
Bangalore, Karnataka 560066
India
Email: avantika.sushilkumar@huawei.com

Xian Zhang
Huawei Technologies
F3-1-B R&D Center, Huawei Base Bantian, Longgang District
Shenzhen, Guangdong 518129
China
Email: zhang.xian@huawei.com

Authors' Addresses

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

Email: dhruv.ietf@gmail.com

Qin Wu
Huawei Technologies
101 Software Avenue, Yuhua District
Nanjing, Jiangsu 210012
China

Email: bill.wu@huawei.com

Vishwas Manral
Nano Sec Co
3350 Thomas Rd.
Santa Clara, CA
United States of America

Email: vishwas@nanosec.io

Zafar Ali
Cisco Systems

Email: zali@cisco.com

Kenji Kumaki
KDDI Corporation

Email: ke-kumaki@kddi.com