Internet Engineering Task Force (IETF)
Request for Comments: 5886
Category: Standards Track
ISSN: 2070-1721

JP. Vasseur, Ed.
Cisco Systems, Inc.
JL. Le Roux
France Telecom
Y. Ikejiri
NTT Communications Corporation

June 2010

A Set of Monitoring Tools for Path Computation Element (PCE)-Based Architecture

Abstract

A Path Computation Element (PCE)-based architecture has been specified for the computation of Traffic Engineering (TE) Label Switched Paths (LSPs) in Multiprotocol Label Switching (MPLS) and Generalized MPLS (GMPLS) networks in the context of single or multiple domains (where a domain refers to a collection of network elements within a common sphere of address management or path computational responsibility such as Interior Gateway Protocol (IGP) areas and Autonomous Systems). Path Computation Clients (PCCs) send computation requests to PCEs, and these may forward the requests to and cooperate with other PCEs forming a "path computation chain".

In PCE-based environments, it is thus critical to monitor the state of the path computation chain for troubleshooting and performance monitoring purposes: liveness of each element (PCE) involved in the PCE chain and detection of potential resource contention states and statistics in terms of path computation times are examples of such metrics of interest. This document specifies procedures and extensions to the Path Computation Element Protocol (PCEP) in order to gather such information.

Status of This Memo

This is an Internet Standards Track document.

This document is a product of the Internet Engineering Task Force (IETF). It represents the consensus of the IETF community. It has received public review and has been approved for publication by the Internet Engineering Steering Group (IESG). Further information on Internet Standards is available in Section 2 of RFC 5741.

Information about the current status of this document, any errata, and how to provide feedback on it may be obtained at http://www.rfc-editor.org/info/rfc5886.

Table of Contents

1. Introduction

The Path Computation Element (PCE)-based architecture has been specified in [RFC4655] for the computation of Traffic Engineering (TE) Label Switched Paths (LSPs) in Multiprotocol Label Switching (MPLS) and Generalized MPLS (GMPLS) networks in the context of single or multiple domains where a domain refers to a collection of network elements within a common sphere of address management or path computational responsibility such Interior Gateway Protocol (IGP) areas and Autonomous Systems.

Path Computation Clients (PCCs) send computation requests to PCEs using PCReq messages, and these may forward the requests to and cooperate with other PCEs forming a "path computation chain". In the case of successful path computation, the computed paths are then provided to the requesting PCC using PCRep messages. The PCReq and PCRep messages are defined in [RFC5440].

In PCE-based environments, it is critical to monitor the state of the path computation chain for troubleshooting and performance monitoring purposes: liveness of each element (PCE) involved in the PCE chain and detection of potential resource contention states and statistics in terms of path computation times are examples of such metrics of interest.

As defined in [RFC4655], there are circumstances in which more than one PCE is involved in the computation of a TE LSP. A typical example is when the PCC requires the computation of a TE LSP where the head-end and the tail-end of the TE LSP do not reside in adjacent domains and there is no single PCE with the visibility of both the head-end and tail-end domain. We call the set of PCEs involved in the computation of a TE LSP a "path computation chain". As further discussed in Section 3.1, the path computation chain may either be static (pre-configured) or dynamically determined during the path computation process.

As discussed in [RFC4655], a TE LSP may be computed by one PCE (referred to as single PCE path computation) or several PCEs (referred to as multiple PCE path computation). In the former case, the PCC may be able to use IGP extensions to check the liveness of the PCE (see [RFC5088] and [RFC5089]) or PCEP using Keepalive messages. In contrast, when multiple PCEs are involved in the path computation chain, an example of which is the Backward Recursive PCE-based Computation (BRPC) procedure defined in [RFC5441], the PCC's visibility may be limited to the first PCE involved in the path computation chain. Thus, it is critical to define mechanisms in order to monitor the state of the path computation chain.

This document specifies PCEP extensions in order to gather various state metrics along the path computation chain. In this document, we call a "state metric" a metric that characterizes a PCE state. For example, such a metric can have a form of a boolean (PCE is alive or not, PCE is congested or not) or a performance metric (path computation time at each PCE).

PCE state metrics can be gathered in two different contexts: in band or out of band. By "in band" we refer to the situation whereby a PCC requests to gather metrics in the context of a path computation request. For example, a PCC may send a path computation request to a PCE and may want to know the processing time of that request in addition to the computed path. Conversely, if the request is "out of band", PCE state metric collection is performed as a standalone request (e.g., check the liveness of a specific path computation chain, collect the average processing time computed over the last 5-minute period on one or more PCEs).

In this document, we define two monitoring request types: general and specific. A general monitoring request relates to the collection of a PCE state metrics that is not coupled to a particular path computation request (e.g., average CPU load on a PCE). Conversely, a specific monitoring request relates to a particular path computation request (processing time to complete the path computation for a TE LSP).

This document specifies procedures and extensions to the Path Computation Element Protocol (PCEP) ([RFC5440]), including new objects and new PCEP messages, in order to monitor the path computation chain and gather various performance metrics.

The message formats in this document are specified using Backus Naur Format (BNF) encoding as specified in [RFC5511].

1.1. Requirements Language

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

2. Terminology

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.

TE LSP: Traffic Engineering Label Switched Path.

3. Path Computation Monitoring Messages

As defined in [RFC5440], a PCEP message consists of a common header followed by a variable-length body made of a set of objects that can be either mandatory or optional. As a reminder, an object is said to be mandatory in a PCEP message when the object must be included for the message to be considered valid. The P flag (defined in [RFC5440]) is located in the common header of each PCEP object and can be set by a PCEP peer to require a PCE to take into account the related information during the path computation. Because the P flag exclusively relates to a path computation request, it MUST be cleared in the two PCEP messages (PCMonReq and PCMonRep message) defined in this document.

For each PCEP message type, a set of rules is defined that specify the set of objects that the message can carry. An implementation MUST form the PCEP messages using the object ordering specified in this document.

In this document, we define two PCEP messages referred to as the Path Computation Monitoring Request (PCMonReq) and Path Computation Monitoring Reply (PCMonRep) messages so as to handle out-of-band monitoring requests. The aim of the PCMonReq message sent by a PCC to a PCE is to gather one or more PCE state metrics on a set of PCEs involved in a path computation chain. The PCMonRep message sent by a PCE to a PCC is used to provide such data.

3.1. Path Computation Monitoring Request (PCMonReq) Message

The Message-Type field of the PCEP common header for the PCMonReq message is set to 8.

There is one mandatory object that MUST be included within a PCMonReq message: the MONITORING object (see Section 4.1). If the MONITORING object is missing, the receiving PCE MUST send a PCErr message with Error-type=6 (Mandatory Object missing) and Error-value=4 (MONITORING object missing). Other objects are optional.

Format of a PCMonReq message (out-of-band request):

```
<PCMonReq Message>::= <Common Header>
                      <MONITORING>
                      <PCC-ID-REQ>
                      [<pce-list>]
                      [<svec-list>]
                      [<request-list>]
```

where:

```
<pce-list>::=<PCE-ID>[<pce-list>]

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

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

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

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

Format of a PCReq message with monitoring data requested (in-band request):

```
<PCReq Message>::= <Common Header>
                   <MONITORING>
                   <PCC-ID-REQ>
                   [<pce-list>]
                   [<svec-list>]
                   <request-list>
```

where:

```
   <pce-list>::=<PCE-ID>[<pce-list>]

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

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

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

where:

<metric-list>::=<METRIC>[<metric-list>]

The SVEC, RP, END-POINTS, LSPA, BANDWIDTH, METRIC, RRO, IRO, and LOAD-BALANCING objects are defined in [RFC5440]. The XRO object is defined in [RFC5521] and the OF object is defined in [RFC5541]. The PCC-ID-REQ object is defined in Section 4.2.

The PCMonReq message is used to gather various PCE state metrics along a path computation chain. The path computation chain may be determined by the PCC (in the form of a series of a series of PCE-ID objects defined in Section 4.3) according to policy specified on the PCC or alternatively may be determined by the path computation procedure. For example, if the BRPC procedure ([RFC5441]) is used to compute an inter-domain TE LSP, the path computation chain may be determined dynamically. In that case, the PCC sends a PCMonReq message that contains the PCEP objects that characterize the TE LSP attributes along with the MONITORING object (see Section 4.1) that lists the set of metrics of interest. If a list of PCEs is present in the monitoring request, it takes precedence over mechanisms used to dynamically determine the path computation chain. If a PCE receives a monitoring request that specifies a next-hop PCE in the PCE list that is unreachable, the request MUST be silently discarded.

Several PCE state metrics may be requested that are specified by a set of objects defined in Section 4. Note that this set of objects may be extended in the future.

As pointed out in [RFC5440], several situations can arise in the form of:

- a bundle of a set of independent and non-synchronized path computation requests,

- a bundle of a set of independent and synchronized path computation requests (SVEC object defined below required), or

- a bundle of a set of dependent and synchronized path computation requests (SVEC object defined below required).

In the case of a bundle of a set of requests, the MONITORING object SHOULD only be present in the first PCReq or PCMonReq message, and the monitoring request applies to all the requests of the bundle, even in the case of dependent and/or synchronized requests sent using more than one PCReq or PCMonReq message.

Examples of requests. For the sake of illustration, consider the three following examples:

Example 1 (out-of-band request): PCC1 makes a request to check the path computation chain that would be used should it request a path computation for a specific TE LSP named T1. A PCMonReq message is sent that contains a MONITORING object specifying a path computation check, along with the appropriate set of objects (e.g., RP, END-POINTS, etc.) that would be included in a PCReq message for T1.

Example 2 (in-band request): PCC1 requests a path computation for a TE LSP and also makes a request to gather the processing time along the path computation chain selected for the computation of T1. A PCReq message is sent that also contains a MONITORING object that specifies the performance metrics of interest.

Example 3 (out-of-band request): PCC2 requests to gather performance metrics along the specific path computation chain <pce1, pce2, pce3, pce7>. A PCMonReq message is sent to PCE1 that contains a MONITORING object and a sequence of PCE-ID objects that identify PCE1, PCE2, PCE3, and PCE7, respectively.

In all of the examples above, a PCRep message (in-band request) or PCMonReq message (out-of-band request) is sent in response to the request that reports the computed metrics.

3.2. Path Monitoring Reply (PCMonRep) Message

The PCMonRep message is used to provide PCE state metrics back to the requester for out-of-band monitoring requests. The Message-Type field of the PCEP common header for the PCMonRep message is set to 9.

There is one mandatory object that MUST be included within a PCMonRep message: the MONITORING object (see Section 4.1). If the MONITORING object is missing, the receiving PCE MUST send a PCErr message with Error-type=6 (Mandatory Object missing) and Error-value=4 (MONITORING object missing).

Other objects are optional.

Format of a PCMonRep (out-of-band request):

```
<PCMonRep Message>::= <Common Header>
                      <MONITORING>
                      <PCC-ID-REQ>
                      [<RP>]
                      [<metric-pce-list>]
```

where:

```
<metric-pce-list>::=<metric-pce>[<metric-pce-list>]

<metric-pce>::=<PCE-ID>
               [<PROC-TIME>]
               [<OVERLOAD>]
```

Format of a PCRep message with monitoring data (in band):

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

where:

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

   <response>::=<RP>
                <MONITORING>
                <PCC-ID-REQ>
               [<NO-PATH>]
               [<attribute-list>]
               [<path-list>]
               [<metric-pce-list>]

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

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

where:

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

 <metric-list>::=<METRIC>[<metric-list>]

 <metric-pce-list>::=<metric-pce>[<metric-pce-list>]

 <metric-pce>::=<PCE-ID>
               [<PROC-TIME>]
               [<OVERLOAD>]
```

The RP and the NO-PATH objects are defined in [RFC5440]. The PCC-ID-REQ object is defined in Section 4.2.

If the path computation chain has been statically specified in the corresponding monitoring request using the series of a series of PCE-ID objects defined in Section 4.3, the monitoring request MUST use the same path computation chain (using the PCE list but in the reverse order).

4. Path Computation Monitoring Objects

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

Several objects are defined in this section that can be carried within the PCEP PCReq or PCRep messages defined in [RFC5440] in the case of in-band monitoring requests (the PCC requests the computation of the TE LSP in addition to gathering PCE state metrics). In the case of out-of-band monitoring requests, the objects defined in this section are carried within PCMonReq and PCMonRep messages.

All TLVs carried in objects defined in this document have the TLV format defined in [RFC5440]:

- Type: 2 bytes
- Length: 2 bytes
- Value: variable

A PCEP object TLV is comprised of 2 bytes for the type, 2 bytes specifying the TLV length, and a value field. The Length field defines the length of the value portion in bytes. The TLV is padded to 4-byte alignment; padding is not included in the Length field (so a 3-byte value would have a length of 3, but the total size of the TLV would be 8 bytes). Unrecognized TLVs MUST be ignored.

4.1. MONITORING Object

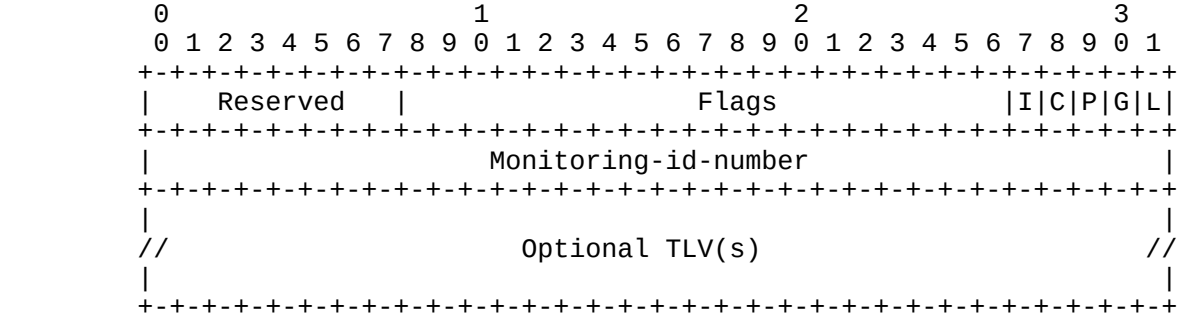
The MONITORING object MUST be present within PCMonReq and PCMonRep messages (out-of-band monitoring requests) and MAY be carried within PCRep and PCReq messages (in-band monitoring requests). There SHOULD NOT be more than one instance of the MONITORING object in a PCMonReq or PCMonRep message: if more than one instance of the MONITORING object is present, the recipient MUST process the first instance and MUST ignore other instances.

The MONITORING object is used to specify the set of requested PCE state metrics.

The MONITORING Object-Class (19) has been assigned by IANA.

The MONITORING Object-Type (1) has been assigned by IANA.

The format of the MONITORING object body is as follows:



Flags: 24 bits

The following flags are currently defined:

L (Liveness) - 1 bit: when set, this indicates that the state metric of interest is the PCE's liveness and thus the PCE MUST include a PCE-ID object in the corresponding reply. The L bit MUST always be ignored in a PCMonRep or PCRep message.

G (General) - 1 bit: when set, this indicates that the monitoring request is a general monitoring request. When the requested performance metric is specific, the G bit MUST be cleared. The G bit MUST always be ignored in a PCMonRep or PCRep message.

P (Processing Time) - 1 bit: the P bit of the MONITORING object carried in a PCMonReq or a PCReq message is set to indicate that the processing times is a metric of interest. If allowed by policy, a PROC-TIME object MUST be inserted in the corresponding PCMonRep or PCRep message. The P bit MUST always be ignored in a PCMonRep or PCRep message.

C (Overload) - 1 bit: The C bit of the MONITORING object carried in a PCMonReq or a PCReq message is set to indicate that the overload status is a metric of interest, in which case an OVERLOAD object MUST be inserted in the corresponding PCMonRep or PCRep message. The C bit MUST always be ignored in a PCMonRep or PCRep message.

I (Incomplete) - 1 bit: If a PCE supports a received PCMonReq message and that message does not trigger any policy violation, but the PCE cannot provide any of the set of requested performance metrics for unspecified reasons, the PCE MUST set the I bit. The I bit has no meaning in a request and SHOULD be ignored on receipt.

Monitoring-id-number (32 bits): The monitoring-id-number value combined with the PCC-REQ-ID identifying the requesting PCC uniquely identifies the monitoring request context. The monitoring-id-number MUST start at a non-zero value and MUST be incremented each time a new monitoring request is sent to a PCE. Each increment SHOULD have a value of 1 and may cause a wrap back to zero. If no reply to a monitoring request is received from the PCE, and the PCC wishes to resend its path computation monitoring request, the same monitoring-id-number MUST be used. Conversely, a different monitoring-id-number MUST be used for different requests sent to a PCE. A PCEP implementation SHOULD checkpoint the Monitoring-id-number of pending monitoring requests in case of restart thus avoiding the reuse of a Monitoring-id-number of an in-process monitoring request.

Unassigned bits are considered as reserved and MUST be set to zero on transmission and ignored on reception.

No optional TLVs are currently defined.

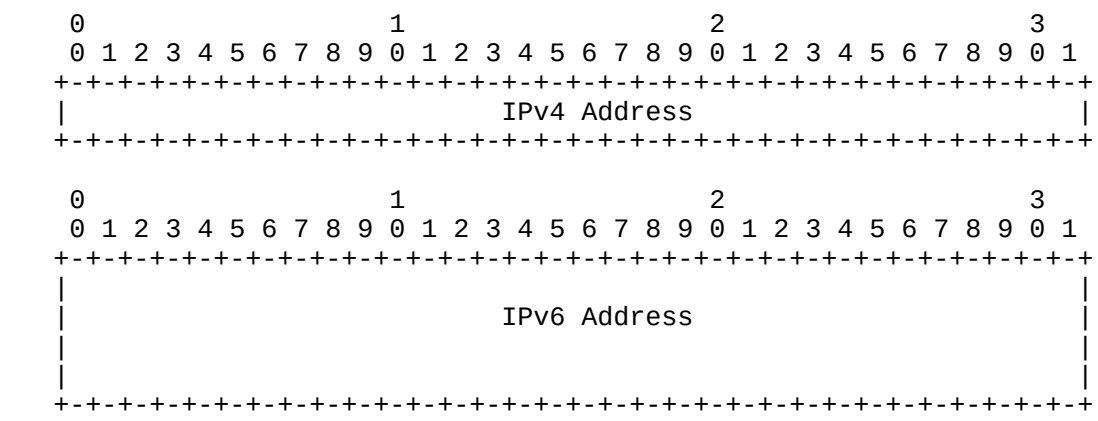
4.2. PCC-ID-REQ Object

The PCC-ID-REQ object is used to specify the IP address of the requesting PCC.

The PCC-ID-REQ MUST be inserted within a PCReq or a PCMonReq message to specify the IP address of the requesting PCC.

Two PCC-ID-REQ objects (for IPv4 and IPv6) are defined. PCC-ID-REQ Object-Class (20) has been assigned by IANA. PCC-ID-REQ Object-Type (1 for IPv4 and 2 for IPv6) has been assigned by IANA.

The format of the PCC-ID-REQ object body for IPv4 and IPv6 are as follows:



The PCC-ID-REQ object body has a fixed length of 4 octets for IPv4 and 16 octets for IPv6.

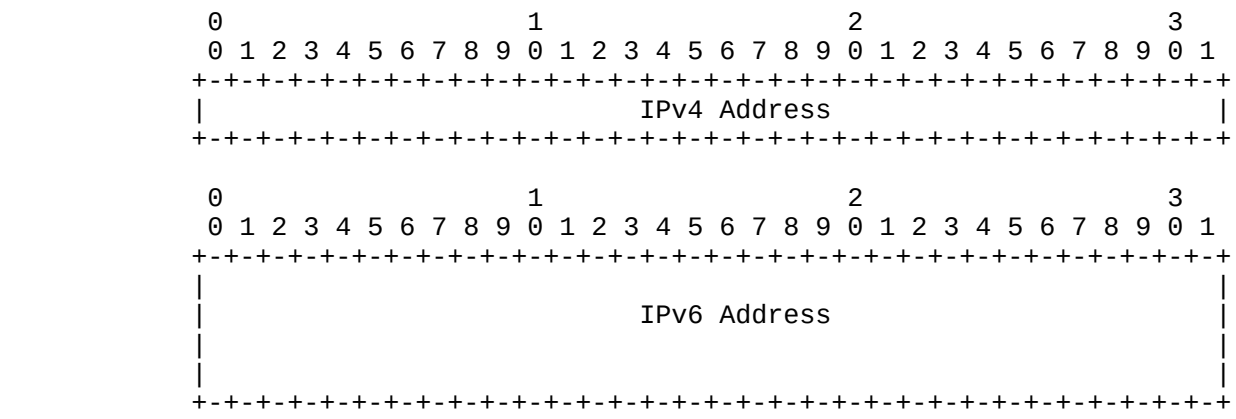
4.3. PCE-ID Object

The PCE-ID object is used to specify a PCE's IP address. The PCE-ID object can either be used to specify the list of PCEs for which monitoring data is requested and to specify the IP address of the requesting PCC.

A set of PCE-ID objects may be inserted within a PCReq or a PCMonReq message to specify the PCE for which PCE state metrics are requested and in a PCMonRep or a PCRep message to record the IP address of the PCE reporting PCE state metrics or that was involved in the path computation chain.

Two PCE-ID objects (for IPv4 and IPv6) are defined. PCE-ID Object-Class (25) has been assigned by IANA. PCE-ID Object-Type (1 for IPv4 and 2 for IPv6) has been assigned by IANA.

The format of the PCE-ID object body for IPv4 and IPv6 are as follows:



The PCE-ID object body has a fixed length of 4 octets for IPv4 and 16 octets for IPv6.

When a dynamic discovery mechanism is used for PCE discovery, a PCE advertises its PCE address in the PCE-ADDRESS sub-TLV defined in [RFC5088] and [RFC5089]. A PCC MUST use this address in PCReq and PCMonReq messages and a PCE MUST also use this address in PCRep and PCMonRep messages.

4.4. PROC-TIME Object

If allowed by policy, the PCE includes a PROC-TIME object within a PCMonRep or a PCRep message if the P bit of the MONITORING object carried within the corresponding PCMonReq or PCReq message is set. The PROC-TIME object is used to report various processing time related metrics.

1) Case of general monitoring requests

   A PCC may request processing time metrics for general monitoring requests (e.g., the PCC may want to know the minimum, maximum, and average processing times on a particular PCE). In this case, general requests can only be made by using PCMonReq/PCMonRep messages. The Current-processing-time field (as explained below) is exclusively used for specific monitoring requests and MUST be cleared for general monitoring requests. The algorithms used by a PCE to compute the minimum, maximum, average, and variance of the processing times are out of the scope of this document (a PCE may decide to compute the minimum processing time over a period of time, for the last N path computation requests, etc.).

2) Case of specific monitoring requests

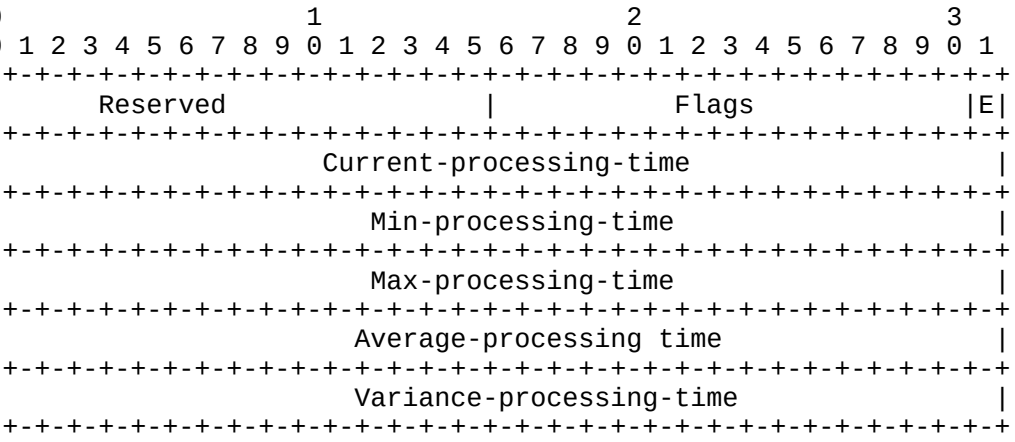
In the case of a specific request, the algorithms used by a PCE to compute the Processing-time metrics are out of the scope of this document, but a flag is specified that is used to indicate to the requester whether the processing time value was estimated or computed. The PCE may either (1) estimate the processing time without performing an actual path computation or (2) effectively perform the computation to report the processing time. In the former case, the E bit of the PROC-TIME object MUST be set. The G bit MUST be cleared and the Min-processing-time, Max-processing-time, Average-processing-time, and Variance-processing-time MUST be set to 0x00000000.

When the processing time is requested in addition to a path computation (case where the MONITORING object is carried within a PCReq message), the PROC-TIME object always reports the actual processing time for that request and thus the E bit MUST be cleared.

The PROC-TIME Object-Class (26) has been assigned by IANA.

The PROC-TIME Object-Type (1) has been assigned by IANA.

The format of the PROC-TIME object body is as follows:



Flags: 16 bits - one flag is currently defined:

E (Estimated) - 1 bit: when set, this indicates that the reported metric value is based on estimated processing time as opposed to actual computations.

Unassigned bits are considered as reserved and MUST be set to zero on transmission.

Current-processing-time: This field indicates, in milliseconds, the processing time for the path computation of interest characterized in the corresponding PCMonReq message.

Min-processing-time: This field indicates, in milliseconds, the minimum processing time.

Max-processing-time: This field indicates, in milliseconds, the maximum processing time.

Average-processing-time: This field indicates, in milliseconds, the average processing time.

Variance-processing-time: This field indicates, in milliseconds, the variance of the processing times.

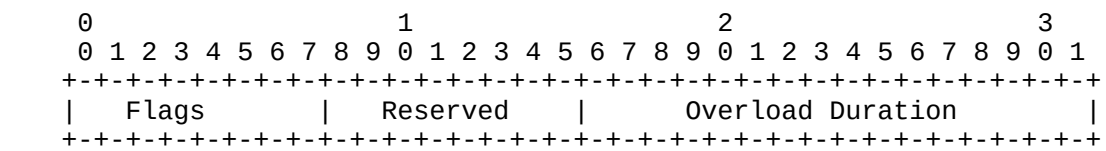
Since the PCC may potentially use monitoring metrics as input to their PCE selection, it MAY be required to normalize how time metrics (along with others metrics described in further revision of this document) are computed to ensure consistency between the monitoring metrics computed by a set of PCEs.

4.5. OVERLOAD Object

The OVERLOAD object is used to report a PCE processing congestion state. Note that "overload" as indicated by this object refers to the processing state of the PCE and its ability to handle new PCEP requests. A PCE is overloaded when it has a backlog of PCEP requests such that it cannot immediately start to process a new request thus leading to waiting times. The overload duration is quantified as being the (estimated) time until the PCE expects to be able to immediately process a new PCEP request.

The OVERLOAD object MUST be present within a PCMonRep or a PCRep message if the C bit of the MONITORING object carried within the corresponding PCMonReq or PCReq message is set and the PCE is experiencing a congested state. The OVERLOAD Object-Class (27) has been assigned by IANA. The overload Object-Type (1) has been assigned by IANA.

The format of the CONGESTION object body is as follows:



Flags: 8 bits - No flag is currently defined.

Overload duration - 16 bits: This field indicates the amount of time, in seconds, that the responding PCE expects that it may continue to be overloaded from the time that the response message was generated. The receiver MAY use this value to decide whether or not to send further requests to the same PCE.

It is worth noting that a PCE along a path computation chain involved in the monitoring request may decide to learn from the overload information received by one of downstream PCEs in the chain.

5. Policy

The receipt of a PCMonReq message may trigger a policy violation on some PCE; in which case, the PCE MUST send a PCErr message with Error-type=5 and Error-value=6.

6. Elements of Procedure

I bit processing: as indicated in Section 4.1, if a PCE supports a received PCMonReq message and that message does not trigger any policy violation, but the PCE cannot provide any of the set of requested performance metrics for unspecified reasons, the PCE MUST set the I bit. Once set, the I bit MUST NOT be changed by a receiving PCE.

Upon receiving a PCMonReq message:

1) As specified in [RFC5440], if the PCE does not support the PCMonReq message, the PCE peer MUST send a PCErr message with Error-value=2 (capability not supported). According to the procedure defined in Section 6.9 of [RFC5440], if a PCC/PCE receives unrecognized messages at a rate equal of greater than specified rate, the PCC/PCE must send a PCEP CLOSE message with close value=5 "Reception of an unacceptable number of unrecognized PCEP messages". In this case, the PCC/PCE must also close the TCP session and must not send any further PCEP messages on the PCEP session.

2) If the PCE supports the PCMonReq message but the monitoring request is prohibited by policy, the PCE must follow the procedure specified in Section 5. As pointed out in Section 4.3, a PCE may still partially satisfy a request, leaving out some of the required data if not allowed by policy.

3) If the PCE supports the PCMonReq and the monitoring request is not prohibited by policy, the receiving PCE MUST first determine whether it is the last PCE of the path computation chain. If the PCE is not the last element of the path computation chain, the PCMonReq message is relayed to the next-hop PCE: such a next hop may be either specified by means of a PCE-ID object present in the PCMonReq message or dynamically determined by means of a procedure outside of the scope of this document. Conversely, if the PCE is the last PCE of the path computation chain, the PCE originates a PCMonRep message that contains the requested objects according to the set of requested PCE states metrics listed in the MONITORING object carried in the corresponding PCMonReq message.

Upon receiving a PCReq message that carries a MONITORING and potentially other monitoring objects (e.g., PCE-ID object):

1) As specified in [RFC5440], if the PCE does not support (in-band) monitoring, the PCE peer MUST send a PCErr message with Error-value=2 (capability not supported). According to the procedure defined in Section 6.9 of [RFC5440], if a PCC/PCE receives unrecognized messages at a rate equal or greater than a specified rate, the PCC/PCE must send a PCEP CLOSE message with close value=5 "Reception of an unacceptable number of unrecognized PCEP messages". In this case, the PCC/PCE must also close the TCP session and must not send any further PCEP messages on the PCEP session.

2) If the PCE supports the monitoring request but the monitoring request is prohibited by policy, the PCE must follow the procedure specified in Section 5. As pointed out in Section 4.3, a PCE may still partially satisfy a request, leaving out some of the required data if not allowed by policy.

3) If the PCE supports the monitoring request and that request is not prohibited by policy, the receiving PCE MUST first determine whether it is the last PCE of the path computation chain. If the PCE is not the last element of the path computation chain, the PCReq message (with the MONITORING object and potentially other monitoring objects such as the PCE-ID) is relayed to the next-hop PCE: such a next hop may be either specified by means of a PCE-ID object present in the PCReq message or dynamically determined by means of a procedure outside of the scope of this document.

Conversely, if the PCE is the last PCE of the path computation chain, the PCE originates a PCRep message that contains the requested objects according to the set of requested PCE states metrics listed in the MONITORING and potentially other monitoring objects carried in the corresponding PCReq message.

Upon receiving a PCMonRep message, the PCE processes the request, adds the relevant objects to the PCMonRep message and forwards the PCMonRep message to the upstream requesting PCE or PCC.

Upon receiving a PCRep message that carries monitoring data, the message is processed, additional monitoring data is added according to this specification, and the message is forwarded upstream to the requesting PCE or PCC.

7. Manageability Considerations

7.1. Control of Function and Policy

It MUST be possible to configure the activation/deactivation of PCEP monitoring on a PCEP speaker. In addition to the parameters already listed in Section 8.1 of [RFC5440], a PCEP implementation SHOULD allow configuring on a PCE whether or not specific, generic, in-band and out-of-band monitoring requests are allowed. Also, a PCEP implementation SHOULD allow configuring on a PCE a list of authorized state metrics (aliveness, overload, processing time, etc.). This may apply to any session in which the PCEP speaker participates, to a specific session with a given PCEP peer or to a specific group of sessions with a specific group of PCEP peers, for instance, the PCEP peers of a neighbor AS.

7.2. Information and Data Models

A new MIB Module may be defined that provides local PCE state metrics, as well as state metrics of other PCEs gathered using mechanisms defined in this document.

7.3. Liveness Detection and Monitoring

This document provides mechanisms to monitor the liveliness and performances of a given path computation chain.

7.4. Verify Correct Operations

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

7.5. Requirements on Other Protocols

Mechanisms defined in this document do not imply any requirements on other protocols in addition to those already listed in [RFC5440].

7.6. Impact on Network Operations

The frequency of PCMonReq messages may impact the operations of PCEs. An implementation SHOULD allow a limit to be placed on the rate of PCMonReq messages sent by a PCEP speaker and processed from a peer. It SHOULD also allow sending a notification when a rate threshold is reached. An implementation SHOULD allow handling PCReq messages with a higher priority than PCMonReq messages. An implementation SHOULD allow the configuration of a second limit for the PCReq message requesting monitoring data.

8. Guidelines to Avoid Overload Thrashing

An important concern while processing overload information is to prevent the overload condition on one PCE simply being moved to another PCE. Indeed, there is a risk that the reaction to an indication of overload will act to increase the amount of overload within the network. Furthermore, this may lead to oscillations between PCEs if the overload information is not handled properly.

This section presents some brief guidance on how a PCC (which term includes a PCE making requests of another PCE) should react when it receives an indication that a PCE is overloaded.

When an overload indication is received (on a PCRep message or on a PCMonRep message), it identifies that new PCReq messages sent to the PCE might be subject to a delay equal to the value in the Overload Duration field (when present).

It also indicates that PCReq messages already queued at the PCE might be subject to a delay. The PCC must decide how to handle new PCReq messages and what to do about PCReq messages already queued at the PCE.

It is RECOMMENDED that a PCC does not cancel a queued PCReq and reissue it to another PCE because of the PCE being overloaded.

Such behavior is likely to result in overload thrashing as multiple PCCs move the PCE queue to another PCE. This would simply introduce additional delay in the processing of all requests. A PCC MAY choose to cancel a queued PCE request if it is willing to sacrifice the request, maybe reissuing it later (after the overload condition has been determined to have cleared by use of a PCMonReq/Rep exchange).

It is then RECOMMENDED to send the cancellation request with a higher priority in order for the overloaded PCE to detect the request cancellation before processing the related request.

A PCC that is aware of PCE overload at one PCE MAY select a different PCE to service its next PCReq message. In doing so, it is RECOMMENDED that the PCC consider whether the other PCE is overloaded or might be likely to become overloaded by other PCCs similarly directing new PCReq messages.

Furthermore, should the second PCE be also overloaded, it is RECOMMENDED not to make any attempt to switch back to the other PCE without knowing that the first PCE is no longer overloaded.

9. IANA Considerations

9.1. New PCEP Message

Each PCEP message has a message type value.

Two new PCEP (specified in [RFC5440]) messages are defined in this document:

Value	Description	Reference
8	Path Computation Monitoring Request (PCMonReq)	This document
9	Path Computation Monitoring Reply (PCMonRep)	This document

9.2. New PCEP Objects

Each PCEP object has an Object-Class and an Object-Type. The following new PCEP objects are defined in this document:

Object-Class Value	Name	Object-Type	Reference
19	MONITORING	1	This document
20	PCC-REQ-ID	1: IPv4 addresses 2: IPv6 addresses	This document
25	PCE-ID	1: IPv4 addresses 2: IPv6 addresses	This document This document
26	PROC-TIME	1	This document
27	OVERLOAD	1: overload	This document

9.3. New Error-Values

A registry was created for the Error-type and Error-value of the PCEP Error Object.

A new Error-value for the PCErr message Error-type=5 (Policy Violation) (see [RFC5440]) is defined in this document.

Error-type	Meaning	Error-value	Reference
5	Policy violation	6: Monitoring message supported but rejected due to policy violation	This document

A new Error-value for the PCErr message Error-type=6 (Mandatory object missing) (see [RFC5440]) is defined in this document.

Error-type	Meaning	Error-value	Reference
6	Mandatory Object missing	4: MONITORING object missing	This document

9.4. MONITORING Object Flag Field

IANA has created a registry to manage the Flag field of the MONITORING object.

New bit numbers may be allocated only by an IETF Review. Each bit should be tracked with the following qualities:

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

Several bits are defined for the MONITORING Object flag field in this document:

Codespace of the Flag field (MONITORING Object)

Bit	Description	Reference
0-18	Unassigned	
19	Incomplete	This document
20	Overload	This document
21	Processing Time	This document
22	General	This document
23	Liveness	This document

9.5. PROC-TIME Object Flag Field

IANA has created a registry to manage the Flag field of the PROC-TIME object.

New bit numbers may be allocated only by an IETF Review. Each bit should be tracked with the following qualities:

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

One bit is defined for the PROC-TIME Object flag field in this document:

Codespace of the Flag field (PROC-TIME Object)

Bit	Description	Reference
0-14	Unassigned	
15	Estimated	This document

9.6. OVERLOAD Object Flag Field

IANA has created a registry to manage the Flag field of the OVERLOAD object.

New bit numbers may be allocated only by an IETF Review. Each bit should be tracked with the following qualities:

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

No Flag is currently defined for the OVERLOAD Object flag field in this document.

Codespace of the Flag field (OVERLOAD Object)

Bit	Description	Reference
0-7	Unassigned	

10. Security Considerations

The use of monitoring data can be used for various attacks such as denial-of-service (DoS) attacks (for example, by setting the C bit and overload duration field of the OVERLOAD object to stop PCCs from

using a PCE). Thus, it is recommended to make use of the security mechanisms discussed in [RFC5440] to secure a PCEP session (authenticity, integrity, privacy, and DoS protection, etc.) to secure the PCMonReq and PCMonRep messages and PCE state metric objects defined in this document. An implementation SHOULD allow limiting the rate at which PCMonReq or PCReq messages carrying monitoring requests received from a specific peer are processed (input shaping) as discussed in Section 10.7.2 of [RFC5440], or from another domain (see also Section 7.6).

11. Acknowledgments

The authors would like to thank Eiji Oki, Mach Chen, Fabien Verhaeghe, Dimitri Papadimitriou, and Francis Dupont for their useful comments. Special thanks to Adrian Farrel for his detailed review.

12. References

12.1. Normative References

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.

[RFC5440] Vasseur, JP., Ed., and JL. Le Roux, Ed., "Path Computation Element (PCE) Communication Protocol (PCEP)", RFC 5440, March 2009.

[RFC5511] Farrel, A., "Routing Backus-Naur Form (RBNF): A Syntax Used to Form Encoding Rules in Various Routing Protocol Specifications", RFC 5511, April 2009.

[RFC5521] Oki, E., Takeda, T., and A. Farrel, "Extensions to the Path Computation Element Communication Protocol (PCEP) for Route Exclusions", RFC 5521, April 2009.

[RFC5541] Le Roux, JL., Vasseur, JP., and Y. Lee, "Encoding of Objective Functions in the Path Computation Element Communication Protocol (PCEP)", RFC 5541, June 2009.

12.2. Informative References

[RFC4655] Farrel, A., Vasseur, J.-P., and J. Ash, "A Path Computation Element (PCE)-Based Architecture", RFC 4655, August 2006.

[RFC5088] Le Roux, JL., Ed., Vasseur, JP., Ed., Ikejiri, Y., and R. Zhang, "OSPF Protocol Extensions for Path Computation Element (PCE) Discovery", RFC 5088, January 2008.

[RFC5089] Le Roux, JL., Ed., Vasseur, JP., Ed., Ikejiri, Y., and R. Zhang, "IS-IS Protocol Extensions for Path Computation Element (PCE) Discovery", RFC 5089, January 2008.

[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, April 2009.

Authors' Addresses

JP. Vasseur (editor)
Cisco Systems, Inc.
1414 Massachusetts Avenue
Boxborough, MA 01719
USA

EMail: jpv@cisco.com

JL. Le Roux
France Telecom
2, Avenue Pierre-Marzin
Lannion 22307
France

EMail: jeanlouis.leroux@orange-ftgroup.com

Yuichi Ikejiri
NTT Communications Corporation
1-1-6, Uchisaiwai-cho, Chiyoda-ku
Tokyo 100-8019
Japan

EMail: y.ikejiri@ntt.com