TEAS Working Group   Y. Lee (Editor)
                                                    Dhruv Dhody (Editor)
Internet Draft   Huawei
Intended Status: Informational
Expires: March 2017   Daniele Ceccarelli
                                                                Ericsson

September 14, 2016

Architecture and Requirement for Distribution of Link-State and TE Information via PCEP.

draft-leedhody-teas-pcep-ls-03

Abstract

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

This document provides possible architectural alternatives for link-state and TE information distribution and their potential impacts on PCE, network nodes, routing protocols etc.

Copyright Notice

Table of Contents

1. Introduction

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

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

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

[RFC7399] touches upon this issue: "It has also been proposed that the PCE Communication Protocol (PCEP) [RFC5440] could be extended to serve as an information collection protocol to supply information from network devices to a PCE. The logic is that the network devices may already speak PCEP and so the protocol could easily be used to report details about the resources and state in the network, including the LSP state discussed in Sections 14 and 15."

[Stateful-PCE] describes a set of extensions to PCEP to provide stateful control. A stateful PCE has access to not only the information carried by the network's Interior Gateway Protocol (IGP), but also the set of active paths and their reserved resources for its computations. PCC can delegate the rights to modify the LSP parameters to an Active Stateful PCE. This requires PCE to quickly be updated on any changes in the Topology and TEDB, so that PCE can meet the need for updating LSPs effectively and in a timely manner.

The fastest way for a PCE to be updated on TED changes is via a direct interface with each network node and with incremental update from each network node.

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

This document proposes alternative architecture approaches for learning and maintaining the Link State (and TE) information directly on a PCE from network nodes as an alternative to IGPs and BGP transport and investigate the impact from the PCE, routing protocol, and network node perspectives.

2. Applicability

Recent development of a stateful PCE Model changes the PCE operation from path computation alone to include the support of PCE-initiated LSPs. With a stateful PCE model, it is also noted that LSP-DB is maintained by the PCE. For LSP state synchronization of stateful PCEs in GMPLS networks, the LSP attributes, such as its bandwidth, associated route as well as protection information etc, should be updated by PCCs to PCE LSP database (LSP-DB) [S-PCE-GMPLS]. To support all these recent changes in a stateful PCE model, a direct PCE interface to each PCC has to be supported. Relevant TE resource and sate information can also be transported from each node to PCE using this PCC-PCE interface via PCEP. Any resource changes in the node and links can also be quickly updated to PCE using this interface. Convergence time of IGP in GMPLS networks may not be quick enough to support on-line dynamic connectivity required for some applications.

New application areas for GMPLS and PCE in optical transport networks include Wavelength Switched Optical Networking (WSON) and Optical Transport Networks (OTN). WSON scenarios can be divided into routing wavelength assignment (RWA) problems where a PCE requires detailed information about switching node asymmetries and wavelength constraints as well as detailed up to date information on wavelength usage per link [RFC6163]. As more data is anticipated to be made available to PCE with addition of OTN and Flex-grid and possible with some optical impairment data even with the minimum set specified in [G.680], the total amount of data requires significantly more information to be held in the TED than is required for other traffic engineered networks. Related to this issue published by [HWANG] indicated that long convergence time and large number of LSAs flooded in the network might cause scalability problems in OSPF-TE and impose limitations on OSPF-TE applications.

There are two main applicability of this alternative proposed by this draft:

- o Where there is a need for a faster incremental link-state and TE resource and state population and convergence at the stateful PCE.
    - . Where there is no IGP or BGP-LS running in the network nodes.
    - . Where there is IGP or BGP-LS running but with a need for a faster incremental link-state (and TE) resource and state population and convergence at the PCE.
        - . A PCE may receive partial Link-state (and TE) resource and state information (say basic TE) from IGP-TE and other information (optical and impairment) from PCEP.
        - . A PCE may receive full Link-state (and TE) resource and state information from both IGP-TE and PCEP.
- o Where there is no IGP or BGP-LS running at the PCE to learn Link-state (and TE) resource and state information.

A PCC may further choose to send only local TE resource and state information or both local and remote learned TE resource and state information. How a PCE manages the TE resource and state information is implementation specific and thus out of scope of this document.

This is also applicable for transporting (abstract) Link-State and TE information from child PCE to a Parent PCE in H-PCE [RFC6805]; as well as for Physical Network Controller (PNC) to Multi-Domain Service Coordinator (MDSC) in Abstraction and Control of TE Networks (ACTN) [ACTN].

This draft does not advocate that the alternative methods specified in this draft should completely replace the IGP-TE as the method of creating the TED. The split between the data to be distributed via an IGP-TE and the information conveyed via one of the alternatives in this document depends on the nature of the network situation. One could potentially choose to have some traffic engineering information distributed via an IGP-TE while other more specialized traffic information is only conveyed to the PCEs via an alternative interface discussed here.

In addition, the methods specified in this draft is only relevant to a set of architecture options where routing decisions are wholly or partially made in the PCE. On the other hand, the networks that do not support IGP-TE/BGP-LS, the method proposed by this draft may be very relevant.

3. Architecture Options

(1) There are two general architectural alternatives based on how nodes get their local link-state (and TE) resource information to the PCEs:

(1.1) All Nodes send local link-state (and TE) resource information to all PCEs;

(1.2) All Nodes send local link-state (and TE) resource information to a designated PCE and have the PCEs share this information with each other.

(2) Further, a designated node (PCC) can share both local and remote link-state (and TE) information to the PCEs, the remote information might be learned at the node via IGP:

(2.1) Designated Node(s) send local and remote link-state (and TE) resource information to all PCEs;

An important functionality that needs to be addressed in each of these approaches is how a new PCE gets initialized in a reasonably timely fashion.

Figures 1-2 show examples of two options for nodes to share local TE resource information with multiple PCEs. As in the IGP case we assume that switching nodes know their local properties and state including the state of all their local links. In these figures the data plane links are shown with the character "o"; Link-state and TE resource information flow from nodes to PCE by the characters "|", "-", "/", or "\"; and PCE to PCE link-state and TE information, if any, by the character "i".

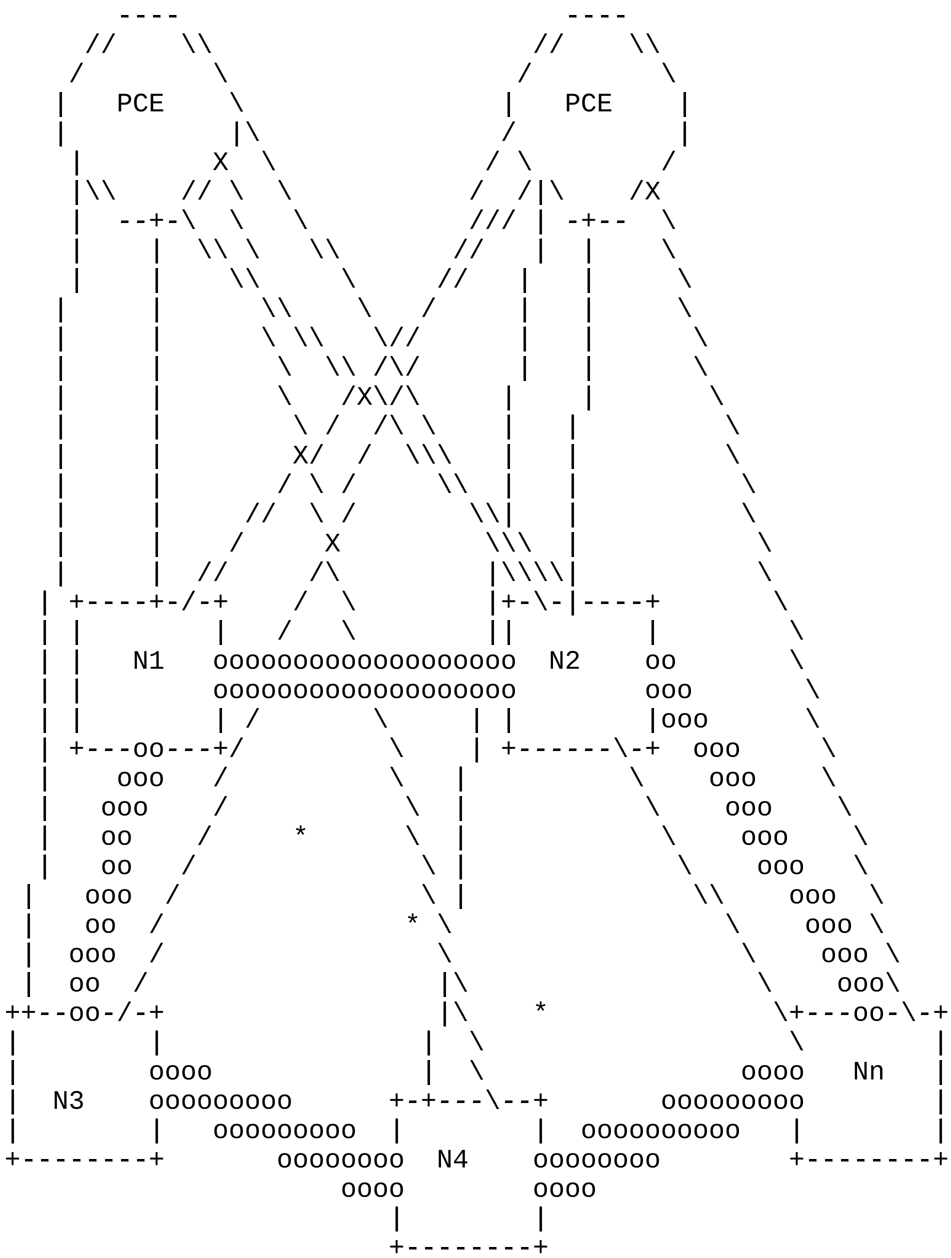

Figure 1 . Nodes send local Link-state and TE information directly to all PCEs

```
                         iiiiiiiiiiiiiiiiii
        iiiiii   ----  iii                 iiiii        ----
       ii    ii//    \\i                       iiiiiiii/    \\
      ii      /        \                             /        \
      i      |   PCE1   |                           |   PCE2   |
     i       |          |                           |          |ii
     i        \        /                             X        /  ii
    i          \\    //                            // \\    //    ii
    i            -//-                             /     --+-       i
    i           //                              //        |        i
   i     +-----/--+                       +----/---+       |        i
   i     |        |                       |        |       |        i
   i     |   N1   ooooooooooooooooooooooooo  N2    oooo    |        i
   i     |        ooooooooooooooooooooooooo        oooo    |        i
   i     |        |                       |        |  oo    |       i
   i     +---oo---+                       +--------+   oo   |        i
   i         oo                                        oo   |        i
   i         oo                                          oo  |        i
   i        oo          *                                oo  |        i
   i       oo                                            oo   |       i
   i       oo                                            oo   |       i
   i       oo                  *                          oo  |       i
   i       oo                                              oo  |      i
   i   +---oo---+                       *               +---oo-+-+    i
   i   |        |                                       |        |    i
   i   |        oooo                                 oooo   Nn   |    i
   i   |  N3    oooooooo       +--------+       ooooooooo        |    i
   ii  |        |    oooooo    |        |  ooooooooooo  |        |   ii
    i  +---\----+       oooooooo  N4    ooooooooo       +--------+   i
    i       \              ooooo        oooo                         i
    ii       \                 |        |                            i
     i        \\               +--------+                           ii
     ii         \              ---                                  i
      ii         \   ----   ---                                     i
       ii         \//    \--                                       i
        ii        /        \                                      ii
          ii     |   PCE3   |                                  iiii
            iiiii|          |                              iiiii
                  \        /                             iii
                   \\    // iiiiiiiii                 iii
                     ----           iiiiiiiiiiiiiiiiiii
```

Figure 2 . Nodes send local Link-state and TE information to one PCE and have the PCEs share TED information

```
          ----                        ----
        /      \                    /      \
       /        \                  /        \
      |   PCE    \                |   PCE    |
      |                         --+          |
       \        /            /     \        /
         \     /            /        \    /
          +-++             /          ----
            |             /
            |            /
            |           /
            |          /
            |         /
            |        /
            |       /
            |      /
            |     /
            |    /
            |   /
            |  /
       +----+-/-+                 +--------+
       |        +                 |        |
       |   N1   ooooooooooooooooooo  N2    oo
       |        ooooooooooooooooooo        ooo
       |        +                 |        |ooo
       +--+oo+--+                 +--------+  ooo
          ooo                                  ooo
         ooo                                    ooo
         oo                                      ooo
         oo                                       ooo
        ooo                                         ooo
        oo                                           oo
       ooo                                            ooo
       oo                                              ooo
   +---oo---+                                       +---oo---+
   |        |                                       |        |
   |        oo                                   oooo   Nn   |
   |  N3    ooooooooo      +--------+       ooooooooo        |
   |        |   ooooooooo  |        |  oooooooooo   |        |
   +--------+       oooooooo  N4    oooooooo        +--------+
                        oooo        oooo
                           |        |
                           +--------+
```

Figure 3. Designated Node sends local and remote Link-state and TE information directly to all PCEs

3.1. Option 1.1: All Nodes Send Local Link-State and TE Info to all PCEs

Architectural alternative 1 shown in Figure 1 illustrates nodes sending their local link-state (and TE) resource information to all PCEs within their domain. As the number of PCEs grows we may have scalability concerns. In particular, each node needs to keep track of which PCE it has sent information to and update that information periodically. However, if we are only talking about 2-3 PCEs, then we do not have this scalability concern.

If a new PCE is added to the domain all nodes must send all its local link-state and TE resource information to that PCE rather than just sending status updates.

3.2. Option 1.2: Each Node Sends Local Link-State and TE Info to one PCE

In this architectural alternative, shown in Figure 2, each node would be associated with one PCE. This implies that each PCE will only have partial link-state (and TE) resource information directly from the nodes. It would be the responsibility of a node to get its local information to its associated PCE, then the PCEs within a domain would then need to share the partial link-state (and TE)resource information they learned from their associated nodes with each other so that they can create and maintain the complete link-state (and TE)resource information.

To allow for this sharing of information PCEs would need to peer with each other. PCE discovery extensions [RFC4674] could be used to allow PCEs to find other PCEs. If a new PCE is added to the domain it would need to peer with at least one other PCE and then PCE synchronization mechanism could then be used to initialize the new PCEs link-state (and TE)resource information.

A number of approaches can be used to ensure control plane resilience in this architecture. (1) Each node can be configured with a primary and a secondary PCE to send its information to; In case of failure of communications with the primary PCE the node would send its information to a secondary PCE (warm standby). (2) Each node could be configured to send its information to two different PCEs (hot standby).

3.3. Option 2.1: Designated Node(s) Send Local and Remote Link-State and TE Info to all PCEs

In this architectural alternative, shown in Figure 3, illustrates designated node(s) sending their local and remote link-state (and TE) resource information to all PCEs within their domain. Designated Node may learn remote information via IGP or BGP-LS. More than one designated node may be used to ensure control plane resilience in this architecture.

3.4. Hierarchy of PCE

Incase of H-PCE [RFC6805], a parent PCE can utilize the mechanism described in this document to prepare the domain topology map by transporting inter-domain link information from child PCE to Parent PCE.

```
                        !!!!!!!!!!!!!!
                        ! Parent PCE !
                        !!!!!!!!!!!!!!
                       / |      |   \
                      /  |      |    \
                     /   |      |     \
                    /    |      |      \
                   /     |       \      \
                  /      |        \     |
                 /       |         \     \
                /        |         |      \
               /         |         |       \
              /     +----|------+  |        \
             /      |  ~~|~~~   |  |         \
            /       |  ~PCE2~   |  |          \
           /        |  ~~~~~~   |  |           \
          /         | Domain 2  |  |            \
         |          |           |  |       +-----\----+
         |          +-----------+  |       |  ~~~~\~  |
        /                          |       |  ~PCE4~  |
   +---/------+                    |       |  ~~~~~~  |
   |  /~~~~~  |                    |       | Domain 4 |
   |  ~PCE1~  |                    |       |          |
   |  ~~~~~~  |                    |       +----------+
   | Domain 1 |                    |
   |          |            +-------|---+
   +----------+            |  ~~~~~|   |
                           |  ~PCE3~   |
                           |  ~~~~~~   |
                           | Domain 3  |
                           |           |
                           +-----------+
```

Figure 4. Child PCE sends some Link-state and TE information to Parent PCE

Further abstracted topology information can be transported from PNC to MDSC in ACTN [ACTN] using this technique described in this document.

3.5. Key Architectural Issues

3.5.1. Nodes Finding PCEs

In all cases, nodes need to send TE information directly to PCEs. Path Computation Clients (PCCs) and network nodes participating in an IGP (with or without TE extensions) have a mechanism to discover a PCE and its capabilities. [RFC4674] outlines the general requirements for this mechanism and extensions have been defined to provide information so that PCCs can obtain key details about available PCEs in OSPF [RFC5088] and in IS-IS [RFC5089].

After finding candidate PCEs, a node would need to see which if any of the PCEs actually want to receive TE information directly from this node.

3.5.2. Node TE Information Update Procedures

First a node must establish an association between itself and a PCE that will be maintaining a link-state and TE information. It is the responsibility of the node to share link-state (and TE) information. This includes local information, e.g., links and node properties or remote information learned from neighbors. General and technology specific information models would specify the content of this information while the specific protocols would determine the format. Note that data plane neighbor information would be passed to the PCE embedded in TE link information.

There will be cases where the node would have to send to the PCE only a subset of TE link information depending on the path computation option. For instance, if the node is responsible for routing while the PCE is responsible for wavelength assignment for the route, the node would only need to send the PCE the WSON link usage information. This path computation option is referred to as separate Fouting (R) and Wavelength Assignment (WA) option in [RFC7449].

3.5.3. PCE Link-state (and TE) Resource Information Maintenance Procedures

The PCE is responsible for creating and maintaining the link-state (and TE) resource information that it will use. Key functions include:

1. Establishing and authenticating communications between the PCE and sources of link-state (and TE) resource information.

2. Timely updates of the link-state (and TE) resource with information received from nodes, peers or other entities.

3. Verifying the validity of link-state (and TE) resource information, i.e., ensure that the network information obtained from nodes or elsewhere is relatively timely, or not stale. By analogy with similar functionality provided by IGPs this can be done via a process where discrete "chunks" of TE resource information are "aged" and discard when expired. This combined with nodes periodically resending their local TE resource information leads to a timely update of TE resource information.

4. Requirements for PCEP extension

The key requirements associated with link-state (and TE) distribution are identified for PCEP and listed in [PCEP-LS].

These new functions required in PCEP to support distribution of link-state (and TE) information are described in [PCEP-LS].

5. New Functions to distribute link-state and TE via PCEP

Several new functions are required in PCEP to support distribution of link-state and TE information. The new functions are:

- Capability advertisement: Advertise capability for link-state and TE information distribution

- Link-State and TE synchronization: Ability to synchronize the Link-state and TE information after session initialization.

o Link-State and TE Report (C-E): a PCC sends a LS and TE report to a PCE whenever the Link-State and TE information changes.

These are listed in some detail in [PCEP-LS]. Also see [PCEP-LS-Optical] for optical extension of [PCEP-LS].

6. Security Considerations

This draft discusses an alternative technique for PCEs to build and maintain a link-state and traffic engineering database. In this approach network nodes would directly send traffic engineering information to a PCE. It may be desirable to protect such information from disclosure to unauthorized parties in addition it may be desirable to protect such communications from interference (modification) since they can be critical to the operation of the network. In particular, this information is the same or similar to that which would be disseminated via a link state routing protocol with traffic engineering extensions.

7. IANA Considerations

This version of this document does not introduce any items for IANA to consider.

8. References

8.1. Normative References

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

[RFC4674] Le Roux, J., Ed., "Requirements for Path Computation Element (PCE) Discovery", RFC 4674, October 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.

[RFC5250] Berger, L., Bryskin, I., Zinin, A., and R. Coltun, "The OSPF Opaque LSA Option", RFC 5250, July 2008.

[RFC5305] Li, T. and H. Smit, "IS-IS Extensions for Traffic Engineering", RFC 5305, October 2008.

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

8.2. Informative References

[JMS] Java Message Service, Version 1.1, April 2002, Sun Microsystems.

[RFC3630] Katz, D., Kompella, K., and D. Yeung, "Traffic Engineering (TE) Extensions to OSPF Version 2", RFC 3630, September 2003.

[RFC4203] Kompella, K., Ed. and Y. Rekhter, Ed., "OSPF Extensions in Support of Generalized Multi-Protocol Label Switching (GMPLS)", RFC 4203, October 2005.

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

[BGP-LS] Gredler, H., Medved, J., Previdi, S., Farrel, A., and S.Ray, "North-Bound Distribution of Link-State and TE information using BGP", draft-ietf-idr-ls-distribution, work in progress.

[S-PCE-GMPLS] X. Zhang, et. al, "Path Computation Element (PCE) Protocol Extensions for Stateful PCE Usage in GMPLS-controlled Networks", draft-ietf-pce-pcep-stateful-pce-gmpls, work in progress.

[RFC7399] A. Farrel and D. king, "Unanswered Questions in the Path Computation Element Architecture", RFC 7399, October 2015.

[RFC7449] Y. Lee, G. Bernstein, "Path Computation Element Communication Protocol (PCEP) Requirements for Wavelength Switched Optical Network (WSON) Routing and Wavelength Assignment", RFC 7449, February 2015.

[RFC4456] Bates, T., Chen, E., and R. Chandra, "BGP Route Reflection: An Alternative to Full Mesh Internal BGP (IBGP)", RFC 4456, April 2006.

[RFC6163] Y. Lee, G. Bernstein, W. Imajuku, "Framework for GMPLS and PCE Control of Wavelength Switched Optical Networks", RFC 6163,

[HWANG] S. Hwang, et al, "An Experimental Analysis on OSPF-TE Convergence Time", Proc. SPIE 7137, Network Architectures, Management, and Applications, November 19, 2008.

[G.680] ITU-T Recommendation G.680, Physical transfer functions of optical network elements, July 2007.

[ACTN] Y. Lee, D. Dhody, S. Belotti, K. Pithewan, and D. Ceccarelli, "Requirements for Abstraction and Control of TE Networks", draft-ietf-teas-actn-requirements, work in progress.

[RFC6805] A. Farrel and D. King, "The Application of the Path Computation Element Architecture to the Determination of a Sequence of Domains in MPLS and GMPLS", RFC 6805, November 2012.

[PCEP-LS] D. Dhody, Y. Lee and D. Ceccarelli, "PCEP Extension for Distribution of Link-State and TE Information", draft-dhodylee-pce-pcep-ls, work in progress.

[PCEP-LS-Optical] Y. Lee and D. Dhody (editors), "PCEP Extension for Distribution of Link-State and TE information for Optical Networks", draft-lee-pce-pcep-ls-optical, work in progress.

[Stateful-PCE] Crabbe, E., Minei, I., Medved, J., and R. Varga, "PCEP Extensions for Stateful PCE", draft-ietf-pce-stateful-pce, work in progress.

[PCE-Initiated] Crabbe, E., Minei, I., Sivabalan, S., and R. Varga, "PCEP Extensions for PCE-initiated LSP Setup in a Stateful PCE Model", draft-ietf-pce-pce-initiated-lsp, work in progress.

Author's Addresses

Young Lee (Editor)
Huawei Technologies
5340 Legacy Drive, Building 3
Plano, TX 75023, USA

Email: leeyoung@huawei.com

Dhruv Dhody (Editor)
Huawei Technologies
Divyashree Technopark, Whitefield
Bangalore, Karnataka 560037
India

EMail: dhruv.ietf@gmail.com

Daniele Ceccarelli
Ericsson
Torshamnsgatan,48
Stockholm
Sweden

EMail: daniele.ceccarelli@ericsson.com

Haomian Zheng
Huawei Technologies Co., Ltd.
F3-1-B R&D Center, Huawei Base,
Bantian, Longgang District
Shenzhen 518129 P.R.China

Email: zhenghaomian@huawei.com

Xian Zhang
Huawei Technologies Co., Ltd.
F3-1-B R&D Center, Huawei Base,
Bantian, Longgang District
Shenzhen 518129 P.R.China

Email: zhangxian@huawei.com