Internet Engineering Task Force (IETF)
Request for Comments: 8051
Category: Informational
ISSN: 2070-1721

X. Zhang, Ed.
Huawei Technologies
I. Minei, Ed.
Google, Inc.

January 2017

Applicability of a Stateful Path Computation Element (PCE)

Abstract

A stateful Path Computation Element (PCE) maintains information about Label Switched Path (LSP) characteristics and resource usage within a network in order to provide traffic-engineering calculations for its associated Path Computation Clients (PCCs). This document describes general considerations for a stateful PCE deployment and examines its applicability and benefits, as well as its challenges and limitations, through a number of use cases. PCE Communication Protocol (PCEP) extensions required for stateful PCE usage are covered in separate documents.

Status of This Memo

This document is not an Internet Standards Track specification; it is published for informational purposes.

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). Not all documents approved by the IESG are a candidate for any level of Internet Standard; see Section 2 of RFC 7841.

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/rfc8051.

Table of Contents

1. Introduction

[RFC4655] defines the architecture for a model based on the Path Computation Element (PCE) for the computation of Multiprotocol Label Switching (MPLS) and Generalized MPLS (GMPLS) Traffic Engineering Label Switched Paths (TE LSPs). To perform such a constrained computation, a PCE stores the network topology (i.e., TE links and nodes) and resource information (i.e., TE attributes) in its TE Database (TED). [RFC5440] describes the Path Computation Element Protocol (PCEP) for interaction between a Path Computation Client (PCC) and a PCE, or between two PCEs, enabling computation of TE LSPs.

As per [RFC4655], a PCE can be either stateful or stateless. A stateful PCE maintains two sets of information for use in path computation. The first is the Traffic Engineering Database (TED), which includes the topology and resource state in the network. This information can be obtained by a stateful PCE using the same mechanisms as a stateless PCE (see [RFC4655]). The second is the LSP State Database (LSP-DB), in which a PCE stores attributes of all active LSPs in the network, such as their paths through the network, bandwidth/resource usage, switching types, and LSP constraints. This state information allows the PCE to compute constrained paths while considering individual LSPs and their inter-dependency. However, this requires reliable state synchronization mechanisms between the PCE and the network, between the PCE and the PCCs, and between cooperating PCEs, with potentially significant control-plane overhead and maintenance of a large amount of state data, as explained in [RFC4655].

This document describes how a stateful PCE can be used to solve various problems for MPLS-TE and GMPLS networks and the benefits it brings to such deployments. Note that alternative solutions relying on stateless PCEs may also be possible for some of these use cases and will be mentioned for completeness where appropriate.

2. Terminology

This document uses the following terms defined in [RFC5440]: PCC, PCE, and PCEP peer.

This document defines the following terms:

Stateful PCE: a PCE that has access to not only the network state, but also to the set of active paths and their reserved resources for its computations. A stateful PCE might also retain information regarding LSPs under construction in order to reduce churn and resource contention. The additional state allows the PCE to compute constrained paths while considering individual LSPs and their interactions. Note that this requires reliable state synchronization mechanisms between the PCE and the network, PCE and PCC, and between cooperating PCEs.

Passive Stateful PCE: a PCE that uses LSP state information learned from PCCs to optimize path computations. It does not actively update LSP state. A PCC maintains synchronization with the PCE.

Active Stateful PCE: a PCE that may issue recommendations to the network. For example, an Active Stateful PCE may use the Delegation mechanism to update LSP parameters in those PCCs that delegate control over their LSPs to the PCE.

Delegation: an operation to grant a PCE temporary rights to modify a subset of LSP parameters on one or more LSPs of a PCC. LSPs are delegated from a PCC to a PCE and are referred to as "delegated" LSPs. The PCC that owns the PCE state for the LSP has the right to delegate it. An LSP is owned by a single PCC at any given point in time. For intra-domain LSPs, this PCC should be the LSP head end.

LSP State Database: information about all LSPs and their attributes.

PCE Initiation: assuming LSP delegation granted by default, a PCE can issue recommendations to the network.

Minimum Cut Set: the minimum set of links for a specific source destination pair that, when removed from the network, results in a specific source being completely isolated from a specific destination. The summed capacity of these links is equivalent to the maximum capacity from the source to the destination by the max-flow min-cut theorem.

3. Application Scenarios

In the following sections, several use cases are described, showcasing scenarios that benefit from the deployment of a stateful PCE.

3.1. Optimization of LSP Placement

The following use cases demonstrate a need for visibility into global LSP states in PCE path computations, and for a PCE control of sequence and timing in altering LSP path characteristics within and across PCEP sessions. Reference topologies for the use cases described later in this section are shown in Figures 1 and 2.

Some of the use cases below are focused on MPLS-TE deployments but may also apply to GMPLS. Unless otherwise cited, use cases assume that all LSPs listed exist at the same LSP priority.

The main benefit in the cases below comes from moving away from an asynchronous PCC-driven mode of operation to a model that allows for central control over LSP computations and maintenance, and focuses specifically on the active stateful PCE model of operation.

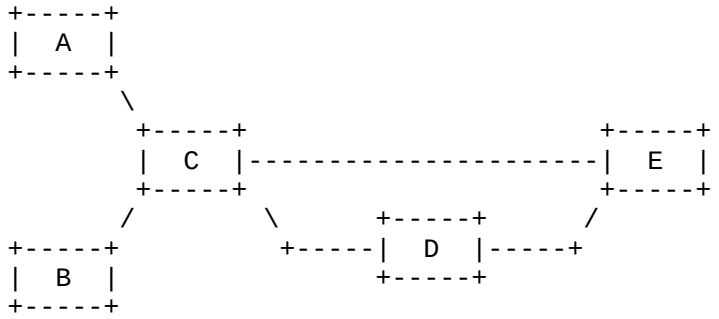

Figure 1: Reference Topology 1

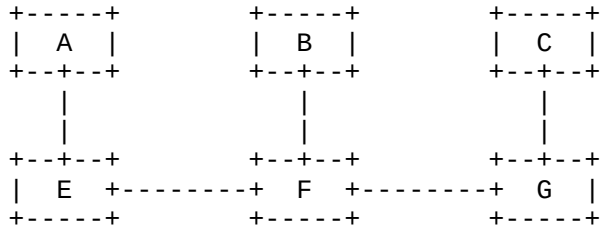

Figure 2: Reference Topology 2

3.1.1. Throughput Maximization and Bin Packing

Because LSP attribute changes in [RFC5440] are driven by Path Computation Request (PCReq) messages under control of a PCC's local timers, the sequence of resource reservation arrivals occurring in the network will be randomized. This, coupled with a lack of global LSP state visibility on the part of a stateless PCE, may result in suboptimal throughput in a given network topology, as will be shown in the example below.

Reference Topology 2 in Figure 2 and Tables 1 and 2 show an example in which throughput is at 50% of optimal as a result of the lack of visibility and synchronized control across PCCs. In this scenario, the decision must be made as to whether to route any portion of the E-G demand, as any demand routed for this source and destination will decrease system throughput.

Link	Metric	Capacity
A-E	1	10
B-F	1	10
C-G	1	10
E-F	1	10
F-G	1	10

Table 1: Link Parameters for Throughput Use Case

Time	LSP	Src	Dst	Demand	Routable	Path
1	1	E	G	10	Yes	E-F-G
2	2	A	B	10	No	---
3	1	F	C	10	No	---

Table 2: Throughput Use Case Demand Time Series

In many cases, throughput maximization becomes a bin-packing problem. While bin packing itself is an NP-hard problem, a number of common heuristics that run in polynomial time can provide significant improvements in throughput over random reservation event distribution, especially when traversing links that are members of the minimum cut set for a large subset of source destination pairs.

Tables 3 and 4 show a simple use case using Reference Topology 1 in Figure 1, where LSP state visibility and control of reservation order across PCCs would result in significant improvement in total throughput.

Link	Metric	Capacity
A-C	1	10
B-C	1	10
C-E	10	5
C-D	1	10
D-E	1	10

Table 3: Link Parameters for Bin-Packing Use Case

Time	LSP	Src	Dst	Demand	Routable	Path
1	1	A	E	5	Yes	A-C-D-E
2	2	B	E	10	No	---

Table 4: Bin-Packing Use Case Demand Time Series

3.1.2. Deadlock

This section discusses the use case of cross-LSP impact under degraded operation. Most existing RSVP-TE implementations will not tear down established LSPs in the event of the failure of the bandwidth increase procedure detailed in [RFC3209]. This behavior is directly implied to be correct in [RFC3209] and is often desirable from an operator’s perspective, because either a) the destination prefixes are not reachable via any means other than MPLS or b) this would result in significant packet loss as demand is shifted to other LSPs in the overlay mesh.

In addition, there are currently few implementations offering dynamic ingress admission control (policing of the traffic volume mapped onto an LSP) at the Label Edge Router (LER). Having ingress admission control on a per-LSP basis is not necessarily desirable from an operational perspective, as a) one must over-provision tunnels significantly in order to avoid deleterious effects resulting from stacked transport and flow control systems (for example, for tunnels that are dynamically resized based on current traffic) and b) there is currently no efficient commonly available northbound interface for dynamic configuration of per-LSP ingress admission control.

Lack of ingress admission control coupled with the behavior in [RFC3209] may result in LSPs operating out of profile for significant periods of time. It is reasonable to expect that these out-of-profile LSPs will be operating in a degraded state and experience traffic loss. Moreover, because those LSPs end up sharing common network interfaces with other LPSs operating within their bandwidth reservations, they will impact the operation of the in-profile LSPs, even when there is unused network capacity elsewhere in the network. Furthermore, this behavior will cause information loss in the TED with regards to the actual available bandwidth on the links used by the out-of-profile LSPs, as the reservations on the links no longer reflect the capacity used.

Reference Topology 1 in Figure 1 and Tables 5 and 6 show a use case that demonstrates this behavior. Two LSPs, LSP 1 and LSP 2, are signaled with demand 2 and routed along paths A-C-D-E and B-C-D-E, respectively. At a later time, the demand of LSP 1 increases to 20. Under such a demand, the LSP cannot be resignaled. However, the existing LSP will not be torn down. In the absence of ingress policing, traffic on LSP 1 will cause degradation for traffic of LSP 2 (due to oversubscription on the links C-D and D-E), as well as information loss in the TED with regard to the actual network state.

The problem could be easily ameliorated by global visibility of the LSP state coupled with PCC-external demand measurements and placement of two LSPs on disjoint links. Note that while the demand of 20 for LSP 1 could never be satisfied in the given topology, isolation from the ill-effects of the (unsatisfiable) increased demand could be achieved.

Link	Metric	Capacity
A-C	1	10
B-C	1	10
C-E	10	5
C-D	1	10
D-E	1	10

Table 5: Link Parameters for the ’Degraded Operation’ Example

Time	LSP	Src	Dst	Demand	Routable	Path
1	1	A	E	2	Yes	A-C-D-E
2	2	B	E	2	Yes	B-C-D-E
3	1	A	E	20	No	---

Table 6: 'Degraded Operation' Demand Time Series

3.1.3. Minimum Perturbation

As a result of both the lack of visibility into the global LSP state and the lack of control over event ordering across PCE sessions, unnecessary perturbations may be introduced into the network by a stateless PCE. Tables 7 and 8 show an example of an unnecessary network perturbation using Reference Topology 1 in Figure 1. In this case, an unimportant (high LSP priority value) LSP (LSP1) is first set up along the shortest path. At time 2, which is assumed to be relatively close to time 1, a second more important (lower LSP-priority value) LSP (LSP2) is established, preempting LSP1 potentially causing traffic loss. LSP1 is then reestablished on the longer A-C-E path.

Link	Metric	Capacity
A-C	1	10
B-C	1	10
C-E	10	10
C-D	1	10
D-E	1	10

Table 7: Link Parameters for the 'Minimum-Perturbation' Example

Time	LSP	Src	Dst	Demand	LSP Prio	Routable	Path
1	1	A	E	7	7	Yes	A-C-D-E
2	2	B	E	7	0	Yes	B-C-D-E
3	1	A	E	7	7	Yes	A-C-E

Table 8: 'Minimum-Perturbation' LSP and Demand Time Series

A stateful PCE can help in this scenario by computing both routes at the same time. The advantages of using a stateful PCE over exploiting a stateless PCE via Global Concurrent Optimization (GCO) are threefold. First is the ability to accommodate concurrent path computation from different PCCs. Second is the reduction of control-plane overhead since the stateful PCE has the route information of the affected LSPs. Thirdly, the stateful PCE can use the LSP-DB to further optimize the placement of LSPs. This will ensure placement of the more important LSP along the shortest path, avoiding the setup and subsequent preemption of the lower priority LSP. Similarly, when a new higher priority LSP that requires preemption of an existing lower priority LSP(s), a stateful PCE can determine the minimum number of lower priority LSPs to reroute using the Make-Before-Break (MBB) mechanism without disrupting any service and then set up the higher priority LSP.

3.1.4. Predictability

Randomization of reservation events caused by lack of control over event ordering across PCE sessions results in poor predictability in LSP routing. An offline system applying a consistent optimization method will produce predictable results to within either the boundary of forecast error (when reservations are over-provisioned by reasonable margins) or to the variability of the signal and the forecast error (when applying some hysteresis in order to minimize churn). Predictable results are valuable for being able to simulate the network and reliably test it under various scenarios, especially under various failure modes and planned maintenances when predictable path characteristics are desired under contention for network resources.

Reference Topology 1 and Tables 9, 10, and 11 show the impact of event ordering and predictability of LSP routing.

Link	Metric	Capacity
A-C	1	10
B-C	1	10
C-E	1	10
C-D	1	10
D-E	1	10

Table 9: Link Parameters for the ’Predictability’ Example

Time	LSP	Src	Dst	Demand	Routable	Path
1	1	A	E	7	Yes	A-C-E
2	2	B	E	7	Yes	B-C-D-E

Table 10: 'Predictability' LSP and Demand Time Series 1

Time	LSP	Src	Dst	Demand	Routable	Path
1	2	B	E	7	Yes	B-C-E
2	1	A	E	7	Yes	A-C-D-E

Table 11: 'Predictability' LSP and Demand Time Series 2

As can be shown in the example, both LSPs are routed in both cases, but along very different paths. This would be a challenge if reliable simulation of the network is attempted. An active stateful PCE can solve this through control over LSP ordering. Based on triggers such as a failure or an optimization trigger, the PCE can order the computations and path setup in a deterministic way.

3.2. Auto-Bandwidth Adjustment

The bandwidth requirements of LSPs often change over time, requiring LSP resizing. In most implementations available today, the head-end node performs this function by monitoring the actual bandwidth usage, triggering a recomputation and resignaling when a threshold is reached. This operation is referred to as "auto-bandwidth adjustment". The head-end node either recomputes the path locally, or it requests a recomputation from a PCE by sending a PCReq message. In the latter case, the PCE computes a new path and provides the new route suggestion. Upon receiving the reply from the PCE, the PCC resignals the LSP in Shared-Explicit (SE) mode along the newly computed path. With a stateless PCE, the head-end node needs to provide the currently used bandwidth and the route information via path computation request messages. Note that in this scenario, the head-end node is the one that drives the LSP resizing based on local information, and that the difference between using a stateless and a passive stateful PCE is in the level of optimization of the LSP placement as discussed in the previous section.

A more interesting smart bandwidth adjustment case is one where the LSP resizing decision is done by an external entity with access to additional information such as historical trending data, application-

specific information about expected demands or policy information, as well as knowledge of the actual desired flow volumes. In this case, an active stateful PCE provides an advantage in both the computation with knowledge of all LSPs in the domain and in the ability to trigger bandwidth modification of the LSP.

3.3. Bandwidth Scheduling

Bandwidth scheduling allows network operators to reserve resources in advance according to the agreements with their customers and allows them to transmit data with a specified starting time and duration, for example, for a scheduled bulk data replication between data centers.

Traditionally, this can be supported by Network Management System (NMS) operation through path pre-establishment and activation on the agreed starting time. However, this does not provide efficient network usage since the established paths exclude the possibility of being used by other services even when they are not used for undertaking any service. It can also be accomplished through GMPLS protocol extensions by carrying the related request information (e.g., starting time and duration) across the network. Nevertheless, this method inevitably increases the complexity of the signaling and routing process.

A passive stateful PCE can support this application with better efficiency since it can alleviate the burden of processing on network elements. This requires the PCE to maintain the scheduled LSPs and their associated resource usage, as well as the ability of head-ends to trigger signaling for LSP setup/deletion at the correct time. This approach requires coarse time synchronization between PCEs and PCCs. With PCE initiation capability, a PCE can trigger the setup and deletion of scheduled requests in a centralized manner, without modification of existing head-end behaviors, by notifying the PCCs to set up or tear down the paths.

3.4. Recovery

The recovery use cases discussed in the following sections show how leveraging a stateful PCE can simplify the computation of recovery path(s). In particular, two characteristics of a stateful PCE are used: 1) using information stored in the LSP-DB for determining shared protection resources and 2) performing computations with knowledge of all LSPs in a domain.

3.4.1. Protection

If a PCC can specify in a request whether the computation is for a working path or for protection and a PCC can report the resource as a working or protection path, then the following text applies. A PCC can send multiple requests to the PCE, asking for two LSPs, and use them as working and backup paths separately. Either way, the resources bound to backup paths can be shared by different LSPs to improve the overall network efficiency, such as m:n protection or pre-configured shared mesh recovery techniques as specified in [RFC4427]. If resource sharing is supported for LSP protection, the information relating to existing LSPs is required to avoid allocation of shared protection resources to two LSPs that might fail together and cause protection contention issues. A stateless PCE can accommodate this use case by having the PCC pass this information as a constraint in the path computation request. A passive stateful PCE can more easily accommodate this need using the information stored in its LSP-DB. Furthermore, an active stateful PCE can help with (re)optimization of protection resource sharing as well as LSP maintenance operation with less impact on protection resources.

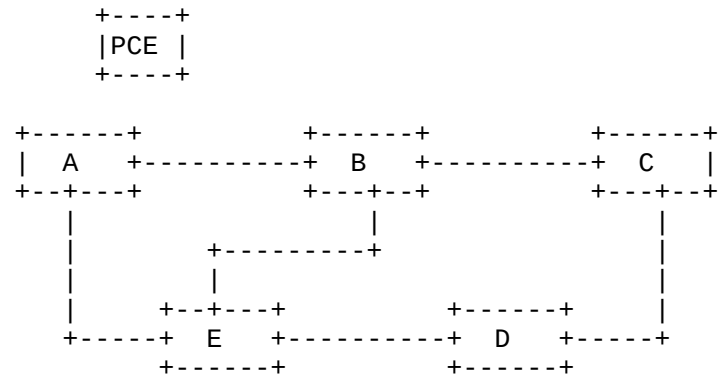

Figure 3: Reference Topology 3

For example, in the network depicted in Figure 3, suppose there exists LSP1 with working path LSP1_working following A->E and with backup path LSP1_backup following A->B->E. A request arrives asking for a working and backup path pair to be computed for LSP2 from B to E. If the PCE decides LSP2_working follows B->A->E, then the backup path LSP2_backup should not share the same protection resource with LSP1 since LSP2 shares part of its resource (specifically A->E) with LSP1 (i.e., these two LSPs are in the same shared risk group). There is no such constraint if B->C->D->E is chosen for LSP2_working.

If a stateless PCE is used, the head node B needs to be aware of the existence of LSPs that share the route of LSP2_working and of the details of their protection resources. B must pass this information to the PCE as a constraint so as to request a path with diversity. Alternatively, a stateless PCE may be able to compute paths diversified by SRLG (Shared Risk Link Group) if TED is extended so that it includes the SRLG information that is protected by a given backup resource, but at the expense of a high complexity in routing. On the other hand, a stateful PCE can get the LSPs information by itself given the LSP identifier(s) and can then find SRLG-diversified protection paths for both LSPs. This is made possible by comparing the LSP resource usage exploiting the LSP-DB accessible by the stateful PCE.

3.4.2. Restoration

In case of a link failure, such as a fiber cut, multiple LSPs may fail at the same time. Thus, the source nodes of the affected LSPs will be informed of the failure by the nodes detecting the failure. These source nodes will send requests to a PCE for rerouting. In order to reuse the resource taken by an existing LSP, the source node can send a PCReq message that includes the Exclude Route Object (XRO) with Fail (F) bit set together with the Record Route Object (RRO) that contains the current route information, as specified in [RFC5521].

If a stateless PCE is used, it might respond to the rerouting requests separately if the requests arrive at different times. Thus, it might result in suboptimal resource usage. Even worse, it might unnecessarily block some of the rerouting requests due to insufficient resources for rerouting messages that arrive later. If a passive stateful PCE is used to fulfill this task, the procedure can be simplified. The PCCs reporting the failures can include LSP identifiers instead of detailed information, and the PCE can find relevant LSP information by inspecting the LSP-DB. Moreover, the PCE can recompute the affected LSPs concurrently while reusing part of the existing LSP’s resources when it is informed of the failed link identifier provided by the first request. This is made possible because the passive stateful PCE can check what other LSPs are affected by the failed link and their route information by inspecting its LSP-DB. As a result, a better performance can be achieved, such as better resource usage or minimal probability of blocking upcoming new rerouting requests sent as a result of the link failure.

If the target is to avoid resource contention within the time window of a high number of LSP rerouting requests, a stateful PCE can retain the under-construction LSP resource usage information for a given time and exclude it from being used for a forthcoming LSP’s request.

In this way, it can ensure that the resource will not be double-booked; thus, the issue of resource contention and computation crank-backs can be alleviated.

3.4.3. SRLG Diversity

An alternative way to achieve efficient resilience is to maintain SRLG disjointness between LSPs, irrespective of whether or not these LSPs share the source and destination nodes. This can be achieved at provisioning time, if the routes of all the LSPs are requested together, using a synchronized computation of the different LSPs with SRLG disjointness constraint. If the LSPs need to be provisioned at different times, the PCC can specify, as constraints to the path computation, a set of SRLGs using the Exclude Route Object [RFC5521]. However, for the latter to be effective, the entity that requests the route to the PCE needs to maintain updated SRLG information regarding all of the LSPs to which it must maintain the disjointness. A stateless PCE can compute an SRLG-disjoint path by inspecting the TED and precluding the links with the same SRLG values specified in the PCReq message sent by a PCC.

A passive stateful PCE maintains the updated SRLG information of the established LSPs in a centralized manner. Therefore, the PCC can specify, as constraints to the path computation, the SRLG disjointness of a set of already established LSPs by only providing the LSP identifiers. Similarly, a passive stateful PCE can also accommodate disjointness using other constraints, such as link, node, or path segment.

3.5. Maintenance of Virtual Network Topology (VNT)

In Multi-Layer Networks (MLN), a Virtual Network Topology (VNT) [RFC5212] consists of a set of one or more TE LSPs in the lower layer, which provides TE links to the upper layer. In [RFC5623], the PCE-based architecture is proposed to support path computation in MLN networks in order to achieve inter-layer TE.

The establishment/teardown of a TE link in VNT needs to take into consideration the state of existing LSPs and/or new LSP request(s) in the higher layer. Hence, when a stateless PCE cannot find the route for a request based on the upper-layer topology information, it does not have enough information to decide whether or not to set up or remove a TE link, which then can result in non-optimal usage of a resource. On the other hand, a passive stateful PCE can make a better decision of when and how to modify the VNT either to accommodate new LSP requests or to reoptimize resource usage across layers irrespective of the PCE models as described in [RFC5623]. Furthermore, given the active capability, the stateful PCE can issue

VNT modification suggestions in order to accommodate path setup requests or reoptimize resource usage across layers.

3.6. LSP Reoptimization

In order to make efficient usage of network resources, it is sometimes desirable to reoptimize one or more LSPs dynamically. In the case of a stateless PCE, in order to optimize network resource usage dynamically through online planning, a PCC must send a request to the PCE together with detailed path/bandwidth information of the LSPs that need to be concurrently optimized. This means that the PCC must be able to determine when and which LSPs should be optimized. In the case of a passive stateful PCE, given the LSP state information in the LSP database, the process of dynamic optimization of network resources can be simplified without requiring the PCC to supply detailed LSP state information. Moreover, an active stateful PCE can even make the process automated by triggering the request. Because a stateful PCE can maintain information for all LSPs that are in the process of being set up and it may have the ability to control timing and sequence of LSP setup/deletion, the optimization procedures can be performed more intelligently and effectively. A stateful PCE can also determine which LSP should be reoptimized based on network events. For example, when an LSP is torn down, its resources are freed. This can trigger the stateful PCE to automatically determine which LSP should be reoptimized so that the recently freed resources may be allocated to it.

A special case of LSP reoptimization is GCO [RFC5557]. Global control of the LSP operation sequence in [RFC5557] is predicated on the use of what is effectively a stateful (or semi-stateful) NMS. The NMS can be either not local to the network nodes, in which case another northbound interface is required for LSP attribute changes, or local/collocated, in which case there are significant issues with efficiency in resource usage. A stateful PCE adds a few features that:

- Roll the NMS visibility into the PCE and remove the requirement for an additional northbound interface.
- Allow the PCE to determine when reoptimization is needed, with which level (GCO or a more incremental optimization).
- Allow the PCE to determine which LSPs should be reoptimized.
- Allow a PCE to control the sequence of events across multiple PCCs, allowing for bulk (and truly global) optimization, LSP shuffling, etc.

3.7. Resource Defragmentation

If LSPs are dynamically allocated and released over time, the resource becomes fragmented. In networks with link bundle, the overall available resource on a (bundle) link might be sufficient for a new LSP request, but if the available resource is not continuous, the request is rejected. Stateful PCEs can be used to perform the defragmentation procedure, because global visibility of LSPs in the network is required to accurately assess resources on the LSPs and to perform defragmentation while ensuring a minimal disruption of the network. This use case cannot be accommodated by a stateless PCE because it does not possess the detailed information of existing LSPs in the network.

Another case of particular interest is the optical spectrum defragmentation in flexible-grid networks. In flexible-grid networks [RFC7698], LSPs with different optical spectrum sizes (such as 12.5GHz, 25GHz, etc.) can coexist so as to accommodate the services with different bandwidth requests. Therefore, even if the overall spectrum size can meet the service request, it may not be usable if the available spectrum resource is not contiguous, but rather fragmented into smaller pieces. Thus, with the help of existing LSP state information, a stateful PCE can make the resource grouped together to be usable. Moreover, a stateful PCE can proactively choose routes for upcoming path requests to reduce the chance of spectrum fragmentation.

3.8. Point-to-Multipoint Applications

PCE has been identified as an appropriate technology for the determination of the paths of Point-to-Multipoint (P2MP) TE LSPs [RFC5671]. The application scenarios and use cases described in Sections 3.1, 3.4, and 3.6 are also applicable to P2MP TE LSPs.

In addition to these, the stateful nature of a PCE simplifies the information conveyed in PCEP messages since it is possible to refer to the LSPs via an identifier. For P2MP, this is an added advantage where the size of the PCEP message is much larger. In case of stateless PCEs, modification of a P2MP tree requires encoding of all leaves along with the paths in a PCReq message. But by using a stateful PCE with P2MP capability, the PCEP message can be used to convey only the modifications (the other information can be retrieved from the identifier via the LSP-DB).

3.9. Impairment-Aware Routing and Wavelength Assignment (IA-RWA)

In Wavelength Switched Optical Networks (WSONs) [RFC6163], a wavelength-switched LSP traverses one or more fiber links. The bit rates of the client signals carried by the wavelength LSPs may be the same or different. Hence, a fiber link may transmit a number of wavelength LSPs with equal or mixed bit-rate signals. For example, a fiber link may multiplex the wavelengths with only 10 Gbit/s signals, mixed 10 Gbit/s and 40 Gbit/s signals, or mixed 40 Gbit/s and 100 Gbit/s signals.

IA-RWA in WSONs refers to the process (i.e., lightpath computation) that takes into account the optical layer/transmission imperfections as additional (i.e., physical layer) constraints. To be more specific, linear and non-linear effects associated with the optical network elements should be incorporated into the route and wavelength assignment procedure. For example, the physical imperfection can result in the interference of two adjacent lightpaths. Thus, a guard band should be reserved between them to alleviate these effects. The width of the guard band between two adjacent wavelengths depends on their characteristics, such as modulation formats and bit rates. Two adjacent wavelengths with different characteristics (e.g., different bit rates) may need a wider guard band and those with the same characteristics may need a narrower guard band. For example, 50 GHz spacing may be acceptable for two adjacent wavelengths with 40 G signals. But for two adjacent wavelengths with different bit rates (e.g., 10 G and 40 G), a larger spacing such as 300 GHz may be needed. Hence, the characteristics (states) of the existing wavelength LSPs should be considered for a new RWA request in WSON.

In summary, when stateful PCEs are used to perform the IA-RWA procedure, they need to know the characteristics of the existing wavelength LSPs. The impairment information relating to existing and to-be-established LSPs can be obtained by nodes in WSON networks via external configuration or other means such as monitoring or estimation based on a vendor-specific impair model. However, WSON-related routing protocols, i.e., [RFC7688] and [RFC7580], only advertise limited information (i.e., availability) of the existing wavelengths, without defining the supported client bit rates. It will incur a substantial amount of control-plane overhead if routing protocols are extended to support dissemination of the new information relevant for the IA-RWA process. In this scenario, stateful PCE(s) would be a more appropriate mechanism to solve this problem. Stateful PCE(s) can exploit impairment information of LSPs stored in LSP-DB to provide accurate RWA calculation.

4. Deployment Considerations

This section discusses general issues with stateful PCE deployments and identifies areas where additional protocol extensions and procedures are needed to address them. Definitions of protocol mechanisms are beyond the scope of this document.

4.1. Multi-PCE Deployments

Stateless and stateful PCEs can coexist in the same network and be in charge of path computation of different types. To solve the problem of distinguishing between the two types of PCEs, either discovery or configuration may be used.

Multiple stateful PCEs can coexist in the same network. These PCEs may provide redundancy for load sharing, resilience, or partitioning of computation features. Regardless of the reason for multiple PCEs, an LSP is only delegated to one of the PCEs at any given point in time. However, an LSP can be redelegated between PCEs, for example, when a PCE fails. [RFC7399] discusses various approaches for synchronizing state among the PCEs when multiple PCEs are used for load sharing or backup and compute LSPs for the same network.

4.2. LSP State Synchronization

The LSP-DB is populated using information received from the PCC. Because the accuracy of the computations depends on the accuracy of the databases used and because the updates must reach the PCE from the network, it is worth noting that the PCE view lags behind the true state of the network. Thus, the use of stateful PCE reduces but cannot eliminate the possibility of crankbacks, nor can it guarantee optimal computations all the time. [RFC7399] discusses these limitations and potential ways to alleviate them.

In case of multiple PCEs with different capabilities coexisting in the same network, such as a passive stateful PCE and an active stateful PCE, it is useful to refer to an LSP, be it delegated or not, by a unique identifier instead of providing detailed information (e.g., route, bandwidth) associated with it, when these PCEs cooperate on path computation, such as for load sharing.

4.3. PCE Survivability

For a stateful PCE, an important issue is to get the LSP state information resynchronized after a restart. LSP state synchronization procedures can be applied equally to a network node or another PCE, allowing multiple ways to reacquire the LSP database on a restart. Because synchronization may also be skipped, if a PCE

implementation has the means to retrieve its database in a different way (for example, from a backup copy stored locally), the state can be restored without further overhead in the network. A hybrid approach where the bulk of the state is recovered locally, and a small amount of state is reacquired from the network, is also possible. Note that locally recovering the state would still require some degree of resynchronization to ensure that the recovered state is indeed up-to-date. Depending on the resynchronization mechanism used, there may be an additional load on the PCE, and there may be a delay in reaching the synchronized state, which may negatively affect survivability. Different resynchronization methods are suited for different deployments and objectives.

5. Security Considerations

This document describes general considerations for a stateful PCE deployment and examines its applicability and benefits, as well as its challenges and limitations through a number of use cases. No new protocol extensions to PCEP are defined in this document.

The PCEP extensions in support of the stateful PCE and the delegation of path control ability can result in more information and control being available for a hypothetical adversary and a number of additional attack surfaces that must be protected. This includes, but is not limited to, the authentication and encryption of PCEP sessions, snooping of the state of the LSPs active in the network, etc. Therefore, documents in which the PCEP protocol extensions are defined need to consider the issues and risks associated with a stateful PCE.

6. References

6.1. Normative References

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

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

[RFC7399] Farrel, A. and D. King, "Unanswered Questions in the Path Computation Element Architecture", RFC 7399, DOI 10.17487/RFC7399, October 2014, <http://www.rfc-editor.org/info/rfc7399>.

6.2. Informative References

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

[RFC4427] Mannie, E., Ed. and D. Papadimitriou, Ed., "Recovery (Protection and Restoration) Terminology for Generalized Multi-Protocol Label Switching (GMPLS)", RFC 4427, DOI 10.17487/RFC4427, March 2006, <http://www.rfc-editor.org/info/rfc4427>.

[RFC5212] Shiomoto, K., Papadimitriou, D., Le Roux, JL., Vigoureux, M., and D. Brungard, "Requirements for GMPLS-Based Multi-Region and Multi-Layer Networks (MRN/MLN)", RFC 5212, DOI 10.17487/RFC5212, July 2008, <http://www.rfc-editor.org/info/rfc5212>.

[RFC5521] Oki, E., Takeda, T., and A. Farrel, "Extensions to the Path Computation Element Communication Protocol (PCEP) for Route Exclusions", RFC 5521, DOI 10.17487/RFC5521, April 2009, <http://www.rfc-editor.org/info/rfc5521>.

[RFC5557] Lee, Y., Le Roux, JL., King, D., and E. Oki, "Path Computation Element Communication Protocol (PCEP) Requirements and Protocol Extensions in Support of Global Concurrent Optimization", RFC 5557, DOI 10.17487/RFC5557, July 2009, <http://www.rfc-editor.org/info/rfc5557>.

[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, <http://www.rfc-editor.org/info/rfc5623>.

[RFC5671] Yasukawa, S. and A. Farrel, Ed., "Applicability of the Path Computation Element (PCE) to Point-to-Multipoint (P2MP) MPLS and GMPLS Traffic Engineering (TE)", RFC 5671, DOI 10.17487/RFC5671, October 2009, <http://www.rfc-editor.org/info/rfc5671>.

[RFC6163] Lee, Y., Ed., Bernstein, G., Ed., and W. Imajuku, "Framework for GMPLS and Path Computation Element (PCE) Control of Wavelength Switched Optical Networks (WSONs)", RFC 6163, DOI 10.17487/RFC6163, April 2011, <http://www.rfc-editor.org/info/rfc6163>.

[RFC7580] Zhang, F., Lee, Y., Han, J., Bernstein, G., and Y. Xu, "OSPF-TE Extensions for General Network Element Constraints", RFC 7580, DOI 10.17487/RFC7580, June 2015, <http://www.rfc-editor.org/info/rfc7580>.

[RFC7688] Lee, Y., Ed. and G. Bernstein, Ed., "GMPLS OSPF Enhancement for Signal and Network Element Compatibility for Wavelength Switched Optical Networks", RFC 7688, DOI 10.17487/RFC7688, November 2015, <http://www.rfc-editor.org/info/rfc7688>.

[RFC7698] Gonzalez de Dios, O., Ed., Casellas, R., Ed., Zhang, F., Fu, X., Ceccarelli, D., and I. Hussain, "Framework and Requirements for GMPLS-Based Control of Flexi-Grid Dense Wavelength Division Multiplexing (DWDM) Networks", RFC 7698, DOI 10.17487/RFC7698, November 2015, <http://www.rfc-editor.org/info/rfc7698>.

Acknowledgements

We would like to thank Cyril Margaria, Adrian Farrel, JP Vasseur, and Ravi Torvi for the useful comments and discussions.

Contributors

The following people all contributed significantly to this document and are listed below in alphabetical order:

Ramon Casellas
CTTC - Centre Tecnologic de Telecomunicacions de Catalunya
Av. Carl Friedrich Gauss n7
Castelldefels, Barcelona 08860
Spain
Email: ramon.casellas@cttc.es

Edward Crabbe
Email: edward.crabbe@gmail.com

Dhruv Dhody
Huawei Technology
Leela Palace
Bangalore, Karnataka 560008
India
Email: dhruv.dhody@huawei.com

Oscar Gonzalez de Dios
Telefonica Investigacion y Desarrollo
Emilio Vargas 6
Madrid, 28045
Spain
Phone: +34 913374013
Email: ogondio@tid.es

Young Lee
Huawei
1700 Alma Drive, Suite 100
Plano, TX 75075
United States of America
Phone: +1 972 509 5599 x2240
Fax: +1 469 229 5397
Email: leeyoung@huawei.com

Jan Medved
Cisco Systems, Inc.
170 West Tasman Dr.
San Jose, CA 95134
United States of America
Email: jmedved@cisco.com

Robert Varga
Pantheon Technologies LLC
Mlynske Nivy 56
Bratislava 821 05
Slovakia
Email: robert.varga@pantheon.sk

Fatai Zhang
Huawei Technologies
F3-5-B R&D Center, Huawei Base
Bantian, Longgang District
Shenzhen 518129
China
Phone: +86-755-28972912
Email: zhangfatai@huawei.com

Xiaobing Zi

Authors' Addresses

Xian Zhang (editor)
Huawei Technologies
F3-5-B R&D Center
Huawei Industrial Base
Bantian, Longgang District
Shenzhen, Guangdong 518129
China

Email: zhang.xian@huawei.com

Ina Minei (editor)
Google, Inc.
1600 Amphitheatre Parkway
Mountain View, CA 94043
United States of America

Email: inaminei@google.com