# Experimental Demonstration of H-PCE with BPG-LS in elastic optical networks

Marta Cuaresma [(1)], Fernando Muñoz del Nuevo [(1)], Sergio Martinez[(1)], Arturo Mayoral [(1)], Oscar González de Dios [(1)], Victor Lopez [(1)], Juan Pedro Fernandez-Palacios[(1)]

[(1)] Telefónica I+D, c/ Don Ramón de la Cruz 84, Madrid, 28006, Spain, Email: jpfpg@tid.es

**Abstract** *Hierarchical PCE is a standard architecture for multi-domain path computation. However, the mechanism to feed the Traffic Engineering Database of the Parent PCE is still under debate. This work validates the use of BGP-LS to build the Parent PCE TEDB and compares two H-PCE algorithms that use different amount of information.*

**Introduction**

The International Engineering Task Force (IETF) has defined the Path Computation Element architecture as a framework capable of solving the path computation problem in complex environments, such as multi-domain transport networks. Mechanisms such as Per-Domain Path Computation, Backward Recursive PCE based Computation (BRPC) and Hierarchical PCE (H-PCE) have been proposed to solve the multi-domain path computation by means of cooperation among different PCEs [1]. The H-PCE architecture is appointed the preferred solution, as it is able to compute optimum end-to-end paths and does not need to have the sequence of domains already fixed.

Current solution draft for the H-PCE [2] is focused on the path computation procedures and the PCEP protocol extensions. However, neither the architecture nor the solution draft define the mechanism that needs to be used to build and populate the parent PCE Traffic Engineering Database (TED). Authors in [3] propose to use PCEP Notifications embedding OSPF-TE Link State Advertisements (LSA) to send the Inter-Domain Link information from child PCEs (cPCEs) to the parent PCE and PCEP Notifications to send reachability information (list of end-points in each domain). This approach has also been experimentally validated in multi-partner testbeds, for a multi-domain WSON scenario in [4] and in a multi-layer multi-domain OBS-WSON scenario in [5]. The main drawbacks of this approach are that it is a non-standard approach and it is not within the scope of PCEP. Furthermore, authors in [6] propose to send aggregated intra-domain topology information in the PCEP notifications. Another approach is to maintain an IGP adjacency between child PCEs and parent PCEs exchanging inter-domain information.

It has been recently propose in the IEFT the North-Bound Distribution of Link-State and TE Information using BGP [7]. This approach is known as BGP-LS and defines a mechanism by which links state and traffic engineering information can be collected from networks and exported to external elements using the BGP routing protocol. In this paper we investigate the use of BGP-LS in the Hierarchical PCE architecture. To validate its use, we have implemented BGP-LS speakers in both child PCE and parent PCE. In particular, this paper is focused on a multi-domain elastic optical network scenario, so extensions for both PCEP and BGP-LS for elastic optical networks are also implemented. BGP-LS allows a fine control of the amount of information sent. This paper studies two H-PCE prodedures that use different amount of information in the parent PCE. In one case, only domain connectivity details are used, and in the other case, BGP-LS is configured to send both intra-domain and inter-domain topologies, giving richer information to the PCE.

The paper is organized as follows. First of all, the H-PCE architecture with BGP-LS is presented. Next, the proposed approach of how to use BGP-LS to exchange information of elastic optical network is explained. Next, the experimental set-up is explained and performance results are shown.

**H-PCE architecture with BPG-LS**

The Hierarchical PCE architecture with BGP-LS is shown in Fig. 1. Each domain has a child PCE (cPCE) that is able to compute paths in the domain. This child PCE has access to a Domain TED, which is built using IGP information. In each domain, a BGP Speaker has access to such domain TED and acts as BGP-LS Route Reflector to provide network topology to the parent PCE (pPCE). Next to the parent PCE, there is a BGP speaker that maintains a BGP session with each of the BGP speakers in the domains to receive the topology and build the parent TED. A policy can be applied to the BGP-LS speakers to decide which information is sent to its peer speaker.

The minimum amount of information that needs

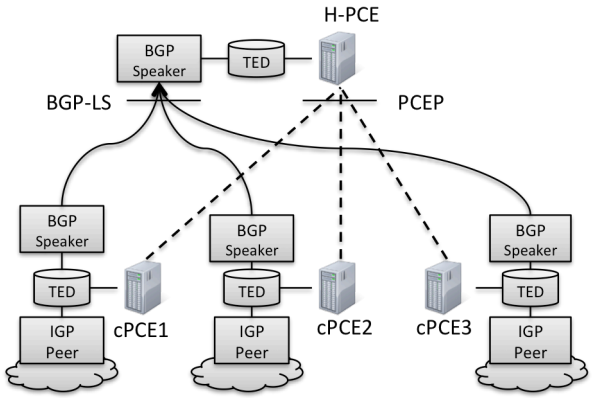


**Fig. 1:** H-PCE architecture with BGP-LS

to be exchanged is the inter-domain connectivity, including the details of the Traffic Engineering Inter-domain Links [1]. With this information, the parent PCE will be able to have access to a domain topology map and its connectivity. Additionally, the BGP-LS speaker can be configured to send the complete list of TE Links, including its details. In this case, the parent PCE will have access to an extended database, with visibility of both intra-domain and inter-domain information and can compute the sequence of domains with better accuracy. Even, the pPCE could have enough information to compute the whole end-to-end path by itself.

**BPG-LS protocol for multi-domain elastic optical networks**

BGP-LS [7] extends the BGP Update messages to advertise link-state topology thanks to new BGP Network Layer Reachability Information (NLRI). In this section we explain how to build the BGP-LS Update messages that contain Inter-domain and intra-domain LSAs. The Link State information is sent in two BGP attributes, the MP_REACH (defined in RFC 4670) and a LINK_STATE attribute (defined in the BGP-LS draft). To describe both the intra and inter domain links, in the MP_REACH attribute, we use a Link NLRI, which contains in the local node descriptors the address of the source, and in the remote descriptors, the address of the destination of the link. The Link Descriptors field has a TLV (Link Local/Remote Identifiers), which carries the prefix of the Unnumbered Interface. In case of the message informs about an intra-domain link, the standard traffic engineering information is included in the LINK_STATE attribute. In addition, the Available Labels TLV [8] is added to the LINK_STATE to include the availability of the frequency slots.

**Experimental validation and performance evaluation**

We have implemented the H-PCE architecture with BGP-LS shown in Fig. 1. The network scenario used for the validation and the performance is a multi-domain elastic optical network. This implemented scenario has three network domains, each with seven network elements and one cPCE. All nodes are virtual machines in a server with two processor Intel Xeon E5-2630 2.30GHz (6 cores each) and 192 GB RAM. Using the netem tool, a delay of 25 ms is added between all network elements (that is between nodes and its cPCE and between cPCE and pPCE). As example of the functional

| | | | | | |
|---|---|---|---|---|---|
| 12 | 26.203094 | 192.168.1.200 | 192.168.1.201 | BGP | OPEN Message |
| 14 | 26.215737 | 192.168.1.201 | 192.168.1.200 | BGP | OPEN Message |
| 20 | 26.260332 | 192.168.1.201 | 192.168.1.200 | BGP | KEEPALIVE Message |
| 21 | 26.264338 | 192.168.1.200 | 192.168.1.201 | BGP | KEEPALIVE Message |
| 164 | 62.572986 | 192.168.2.200 | 192.168.1.201 | BGP | OPEN Message |
| 166 | 62.596196 | 192.168.1.201 | 192.168.2.200 | BGP | OPEN Message |
| 168 | 62.600891 | 192.168.1.201 | 192.168.2.200 | BGP | KEEPALIVE Message |
| 174 | 62.632379 | 192.168.2.200 | 192.168.1.201 | BGP | KEEPALIVE Message |
| 464 | 101.332796 | 192.168.3.200 | 192.168.1.201 | BGP | OPEN Message |
| 465 | 101.346238 | 192.168.1.201 | 192.168.3.200 | BGP | OPEN Message |
| 469 | 101.361512 | 192.168.1.201 | 192.168.3.200 | BGP | KEEPALIVE Message |
| 472 | 101.383922 | 192.168.3.200 | 192.168.1.201 | BGP | KEEPALIVE Message |
| 484 | 102.533076 | 192.168.2.200 | 192.168.1.201 | BGP | UPDATE Message |
| 485 | 102.543051 | 192.168.2.200 | 192.168.1.201 | BGP | UPDATE Message |

**Fig. 2: BGP-LS message exchange**

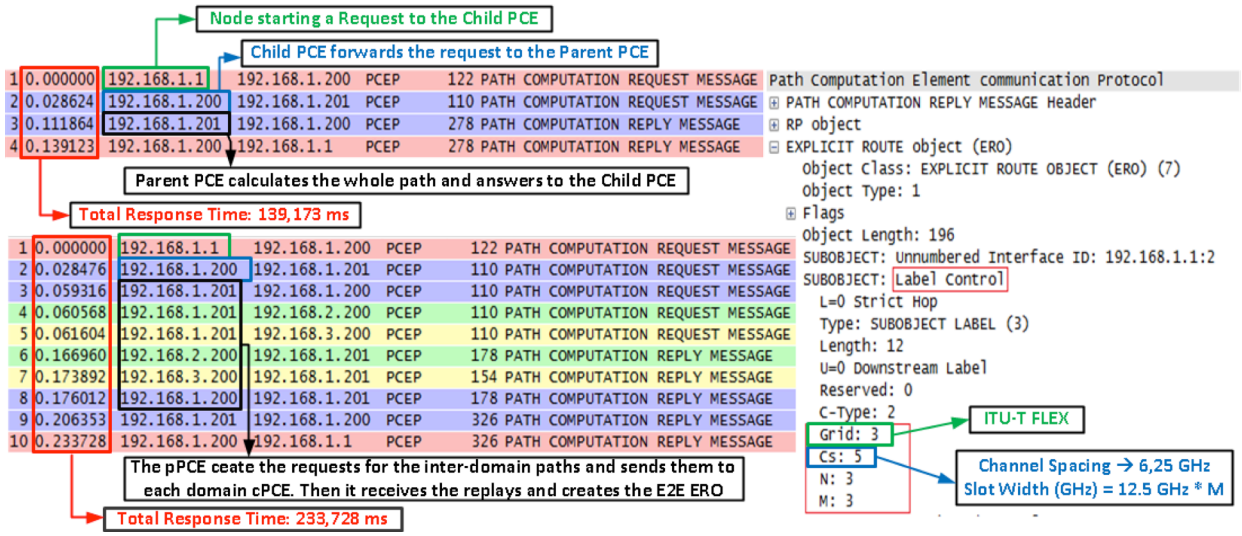


**Fig. 3:** PCEP message flow for a multi-domain request with LMDMTD (top left) and with DMDMTD (bottom left). Detailed ERO object (right)

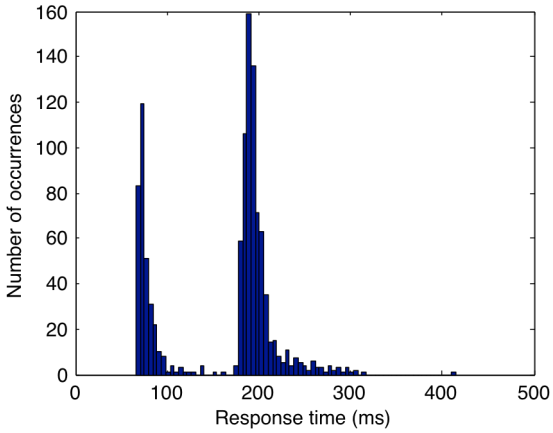


**Fig. 4:** Relative frequency histogram for the response time for DMDMTD algorithm

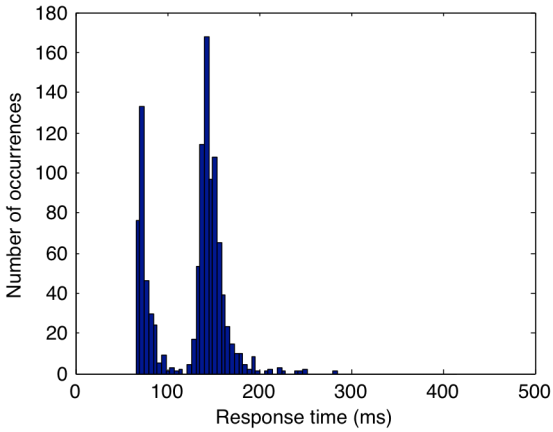


**Fig. 5:** Relative frequency histogram for the response time for LMDMTD algorithm

validation, the Whireshark capture in Fig. 2 depicts the BGP-LS message exchange to feed the pPCE topology/TED and Fig. 3 the PCEP message exchange to request multi-domain paths in H-PCE network architecture for two algorithms, explained below.

The PCEP message flow (Fig. 3) starts with the request from the source transport node to the cPCE asking for a SSON path. Since the destination is not located at the source domain, the cPCE forwards the request using the same parameters. Two algorithms are implemented. In the Local Multi-Domain Minimum Transit Domains (LMDMTD), BGP policy is configured to send all the topology, and the parent PCE computes the shortest domain sequence, and then, for each domain, in parallel, applies a RSA algorithm (Fig. 3, upper left). In the Distributed Multi-Domain Minimum Transit Domains (DMDMTD), the classical H-PCE procedure in which the child PCEs are queried is followed (Fig. 3, lower left). The c PCEs use the same RSA as in the LMDMTD case.

In the performance evaluation, 1000 requests of 100 Gpbs connections are generated between every pair of nodes. Thus, both intra-domain and multi-domain requests are generated. In this paper, computation times are measured for each algorithm (resource reservation time in the control plane is not included) from the time when the node send the PCEP Request until it receives the PCEP Response. Fig. 4 presents the histogram of the total response time for the DMDMTD algorithm. The main source of response time comes from the delays between node and cPCE (50 ms RTT) and between cPCEs and pPCE (50 ms RTT). Fig. 5 shows the histogram of the other procedure of computing multi-domain paths (LMDMTD) in which all the necessary information is provided to the pPCE by BGP-LS. The pPCE computes the whole path itself avoiding the communication with the cPCEs for each intra-domain ERO computation.

The pPCE is able to perform in parallel the domain RSAs without much penalty and needs less time than DMDMTD (Fig. 4) because there is no propagation delay as consequence of the multiple domain requests between the pPCE and the cPCEs.

## Conclusions

The use of BGP-LS in the Hierarchical PCE architecture has been experimentally validated. Two policies in BGP-LS were tested, one sending only inter-domain information and another sending the full topology. When the pPCE used the full topology, a 20% reduction in the computing time, despite the higher complexity of the algorithm, due to the reduction of PCEP Requests to the cPCEs needed.

## Acknowledgements

The research leading to these results has received funding from the European Community's Seventh Framework Programme under grant agreement n° 317999 (IDEALIST project).

## References

[1] D. King and A. Farrel, RFC 6805, Nov 2012
[2] F. Zhang *et al.* draft-zhang-pce-hierarchy-extensions-03, Feb 2013
[3] R. Casellas *et al.*, OThI5, OFC 2011
[4] F. Paoluci *et al.*, OM3G.3, OFC 2012.
[5] O. Gonzalez de Dios *et al.*, OTh1H.2, OFC 2013
[6] R. Casellas *et al.*, “Dynamic Virtual Link Mesh Topology Aggregation in Multi-Domain Translucent WSON with H-PCE”, ECOC 2011
[7] Gredler et al. draft-ietf-idr-ls-distribution-02, Feb 2013
[8] G. Bernstein et al. draft-ietf-ccamp-general-constraint-encode-10, Nov 2012.