Introducing Database Communication Technologies for TED Replication in Multi-domain Networks

(Invited Paper)

G. Cecchetti*, F. Cugini§, F. Paolucci*, P. Castoldi*, A. Giorgetti*

*Scuola Superiore Sant'Anna TeCIP, Pisa, Italy

§CNIT, Pisa, Italy

Abstract—In multi-domain transport networks, exchange of Traffic Engineering information is required to enable effective end-to-end service provisioning and restoration by efficiently utilizing network resources. So far, several solutions have been proposed by the communication community such as the Hierarchical Path Computation Element (H-PCE) architecture. Using the H-PCE architecture a parent PCE is responsible for inter-domain path computation, while a dedicated child PCE performs intra-domain path computation within each domain. However, this approach can introduce scalability concerns especially under dynamic traffic condition such as during restoration because all path computation procedures are coordinated by the parent PCE and may require the exchange of many control messages.

This paper proposes a standard communication among database systems located at the child PCEs, to exchange and share YANG-based Traffic Engineering information in multi-domain networks. By exploiting currently available database technologies, scalable and predictable performance is demonstrated for both replication mechanisms among child PCEs and information retrieval from the stored databases. Thus, this proposal enables the sharing of intra-domain information at each cPCE that can be locally used, upon failure, to speed-up the recovery procedure.

I. INTRODUCTION

Exchange of Traffic Engineering (TE) information is required in multi-domain transport networks to provide constraint-based end-to-end service provisioning while efficiently utilizing network resources [1], [2]. However, including the whole network in a single routing area is not a feasible option for scalability reasons, thus several solutions have been proposed to share the required network resource information among the several domains. The Hierarchical Path Computation Element (H-PCE) architecture, sketched in Fig. 1, proposes a single parent PCE (pPCE) to be responsible for inter-domain path computations, whereas, in each domain, a dedicated child PCE performs intra-domain path computations. Within the H-PCE architecture several methods have been proposed to store, update, and retrieve the intra-domain information at the pPCE, such as the recently standardized Border Gateway Protocol with Link State extensions (BGP-LS [3], [4]). However, this approach still introduces scalability concerns especially under dynamic traffic condition such as during restoration, because all the computation procedures for inter-domain paths directly involve the pPCE. Alternatively, emerging solutions propose to use the BGP-LS protocol to exchange the intra-domain information among peer PCEs [5].
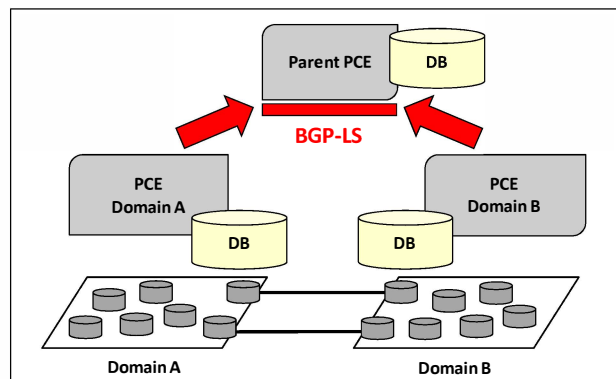

Fig. 1. Traditional H-PCE architecture employing communication protocols (e.g., BGP-LS) for exchange of TE information.

This solution, illustrated in Fig. 2, is interesting for solving the scalability concerns at the pPCE but is still affected by the constraints implied by the utilization of a specific protocol.

Indeed, the traditional way of defining and/or enhancing communication protocols is not free from limitations. First, the whole process is typically slow: years are usually needed to define protocol requirements, protocol specifications, perform related software implementation and run inter-operability tests. Moreover, inter-operability is always challenging to be achieved among multiple different implementations. A second drawback of traditional communication protocols is that the amount of TE information advertised from one domain to another is typically limited by potential scalability issues that may affect the receiving domain. However, the remote domain may not be overloaded and could be effortlessly able to handle all needed information, thus potentially improving path computation efficiency. Third, in traditional approaches (see Fig. 1, and Fig. 2), specific internal procedures need to be implemented at the controller to store and retrieve the information in its local database. In particular, a controller has to run internal API to retrieve TE parameters from its local database and encapsulate them into a communication protocol or into a northbound interface for service use. At the remote controller, TE parameters are de-capsulated and inserted through internal API into its local database. All this operations introduce controller complexity and limit the capacity to rapidly enhance controller capabilities.
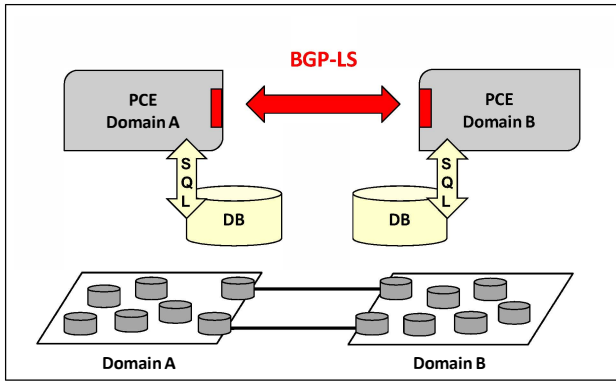
Fig. 2. Emerging architectures employing communication protocols (e.g., BGP-LS) for exchange of TE information.
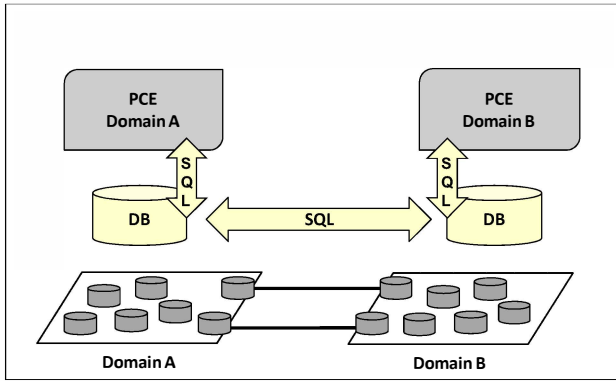

Fig. 3. Proposed solution employing DB synchronization protocols (e.g., SQL) for exchange of YANG-defined TE info.

This paper proposes to exploit the currently available database (DB) technologies to improve the performance in the retrieval and synchronization of TE information. A specific use-case is illustrated, i.e., traffic recovery upon network failure in multi-layer multi-domain networks; where the application of the proposed solution enables the computation of backup paths at the local cPCE, thus alleviating possible scalability issues at the pPCE and reducing the expected recovery time. More specifically, in our previous work [6], we proposed to exploit DB technologies to provide direct access to service API (e.g., for effective correlation of TE information in multi-layer networks). In this paper, we propose to also rely on the DB communication facilities and synchronization protocols as an attractive, already available, alternative with respect to currently available communication protocols. Indeed, in the last years, database technologies have evolved providing impressive performance. The use of such DB communication facilities would significantly speed up the support of newly introduced TE parameters, significantly improve scalability performance and reduce controller operations to handle database parameters.

II. RELATED WORK

Several TE solutions have been proposed for single-domain networks based on the GMPLS/PCE control plane. In such networks the routing protocol (e.g., OSPF-TE) advertises available resources [7], [8]. Therefore effective path computation can be performed either locally or by a dedicated PCE. Conversely, in multi-domain networks resource advertisement were originally based on Border Gateway Protocol (BGP) that only exchanges reachability information not considering resource availability. Therefore, effective inter-domain path computation strategies were inhibited [9]. To address this issue, the PCE architecture has been extended to support inter-domain path computation using a coordinated PCE Protocol (PCEP) communication process among PCEs. As an example, the Backward Recursive Path Computation (BRPC) procedure has been proposed to identify the optimal path considering the current resource utilization [10], [11]. However, those procedures assume that the sequence of domains to be traversed is known in advance and require complex communications among PCEs.

Thus, the H-PCE architecture has been proposed, where the pPCE is introduced to coordinate the inter-domain path computation including the selection of the sequence of domains [12], [13]. Originally, the pPCE has been designed including only inter-domain connectivity information. Later, it has been extended to dynamically ask the cPCEs, for the computation of the several intra-domain path segments. However, also this solution may introduce control plane scalability concerns because of the high number of PCEP messages required between pPCE and cPCEs [2], [14], [15].

Afterwards, new proposals emerged to enable effective inter-domain path computation without requiring extensive communication among PCEs. Specifically, the inclusion of intra-domain information at the pPCE has been proposed [16], using the recently proposed TE Link State Information extension to BGP (BGP-LS), as experimentally demonstrated in [17], [18]. However, this BGP-LS solution is still based on OSPF-TE advertisements and can therefore suffer in dynamic traffic scenarios such as during restoration. As an example, the work in [19] considers dynamic restoration in multi-domain elastic optical networks and demonstrates a blocking probability of several percentage points also with very low traffic load.

With respect to the aforementioned works, the proposal in this paper can be considered as a further step toward fast and effective inter-domain path computation that can be especially useful during traffic recovery. Indeed, the proposed method enables cPCEs to locally store the intra-domain information of other domains. In such a scenario, upon failure occurrence in a domain, the cPCE of each other domain may easily acquire the information regarding the failed resources and operate on the shared database to retrieve the information required to start the recovery procedures, such as the list of disrupted traffic flows. Then, backup paths can be locally computed at the cPCE without requiring complex communication procedures involving pPCE or othe cPCEs.

III. DATABASE TECHNOLOGIES FOR EXCHANGE OF TE INFORMATION

As shown in Fig. 3, this paper proposes to exchange TE information among controllers through standard communication among relational database management systems (RDBMS), operating by means of standard Structured Query Language (SQL). To achieve inter-operability, besides communication protocol and SQL, standard definitions of network and topology parameters are considered. In this work, we rely on the YANG data modelling for network topologies specified in [20] and related documents [21], [22]. The basic YANG description includes abstract and layer/protocol-independent networks and topologies, enabling an inventory of layered networks and nodes. The basic description is then augmented to provide specific YANG definitions for Layer 2 and Layer 3 parameters. Layer 3 description is then further augmented to standardize YANG definitions for specific protocol implementations, like Open Shortest Path First (OSPF) with TE information.

In OSPF-TE, information referred to point-to-point unidirectional links include, besides router and interface IDs for identification, parameters like TE metric, max and unreserved bandwidth, and administrative color. For example, the TE Metric is defined as a leaf *te-metric* of the container *link-tlv* with type *uint32*. Given a standard YANG model to describe network resources, once a controller stores or updates information locally, through standard SQL, the policy-based DB synchronization takes place in an automatic way, without additional operations directly performed by the controller. In particular, DB replication is applied, copying and distributing database data from one database to another, achieving synchronization among databases to maintain consistency.

Several types of data replication technologies can be considered. In this work, we focus on transactional replication. This type typically starts with a snapshot of the database to be replicated. After the initial snapshot, modifications are distributed, in the same order as they occurred, guaranteeing consistency. This type of replication is typically utilized in server-to-server communications to address very high volume of insert, update, and delete activities, applying incremental changes with low latency in data replication. Potential conflicts are resolved by design through the execution of isolated (atomic) transactions.

Transactional replication is suitable for multi-domain networks since each controller is responsible for the resources under its domain of visibility, and it is the only element updating the related DB info (i.e., DB master). No conflicts occur in the update process since only passive replication is applied. Moreover, changes may involve high volume of modifications and low latency is an important requirement to achieve fast network convergence time. Especially, during highly dynamic traffic conditions such as during the recovery procedures required after a network failure. With respect to traditional telecommunication protocols where the case of a single parameter variation (e.g., unreserved link bandwidth) requires the entire set of link parameters to be re-advertised, in the case of DB replication just the specific parameter variation is replicated, i.e., not including the whole set of TE link info. This significantly simplifies the TE exchange process towards remote domains.

Replication occurs at transaction commit time, by broadcasting transaction update to the cluster. Two types of replication can be considered: synchronous and asynchronous. The former guarantees that changes happening on one DB are synchronously propagated to other master DBs of the cluster. Such mechanism is also named as multi-master replication. The latter type provides no guarantees about the delay between applying changes on master DB and the propagation to slave DB. This mechanism is typically named as master-slave replication. Synchronous replication is typically highly available and always consistent. However, it is typically slower than asynchronous replication because the data is considered available only upon the slowest DB is successfully synchronized with all other DBs. Both multi-master and master-slave replication are of interest in the context of multi-domain networks. For example, in the case of multi-domain single-Carrier networks, a multi-master configuration may be preferable for scalability, availability and consistency reasons. On the other hand, in the case of multi-Carrier networks, a master-slave configuration may be more suitable because one domain is publisher (master) of its own resources and subscriber (slave) of other domain resources.

IV. USE-CASE: RECOVERY IN MULTI-LAYER MULTI-DOMAIN NETWORKS

This section describes the utilization of the proposed solution in the use case of a network link failure. Fig. 4 illustrates the considered multi-layer (i.e., optical and IP/MPLS layers) and multi-domain (i.e., two domains) network. It is assumed that the optical layer and the IP/MPLS layer in each domain share a common database as described in [6]. Each network site includes an IP/MPLS router, equipped with Packet Switching Capable (PSC) interfaces; and an optical node (i.e., ROADM) equipped with Lambda Switching Capable (LSC) interfaces. Each router interface is physically attached to a ROADM interface (i.e., transponder). A lightpath is established between each pair of adjacent ROADMs (for clearness, not all lightpaths are illustrated in Fig. 4). Moreover, three additional lightpaths are established: LSP LSC 1 is established between ROADM $A1$ and $A3$ within domain A; LSP LSC 3 is established between ROADM $B4$ and $B2$ within domain B; LSP LSC 4 is established between ROADM $B1$ and $B2$ in within domain B, passing through nodes $B4$ and $B3$. The obtained virtual topology at the IP/MPLS layer is shown in Fig. 5. Then, several PSC LSPs are established at the IP/MPLS layer, including an inter-domain LSP from $a1$ to $b2$ (i.e., LSP PSC 1), and the intra-domain LSP from $b1$ to $b2$ (LSP PSC 2), see Fig. 5.

Fig. 6 reports the relevant entries of the optical layer and IP/MPLS layer tables stored in the database of each domain. The optical layer tables are used to associate the physical cards of each node to the established lightpaths. In turn, the
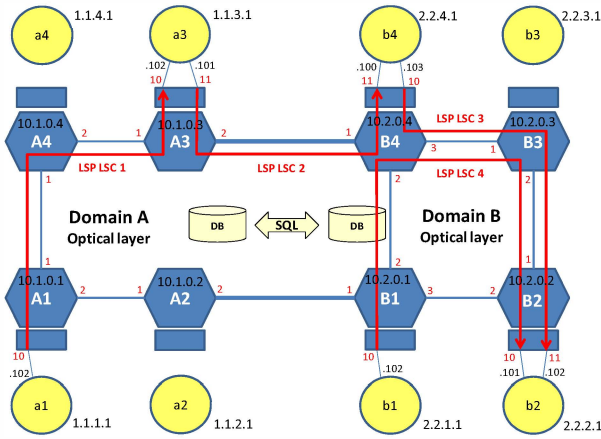
Fig. 4. Optical layer of the test network, the network is composed of 2 domains, each domain is composed of 4 nodes including an optical ROADM and an IP/MPLS router. Four lightpaths (i.e., LSP LSC) are illustrated.
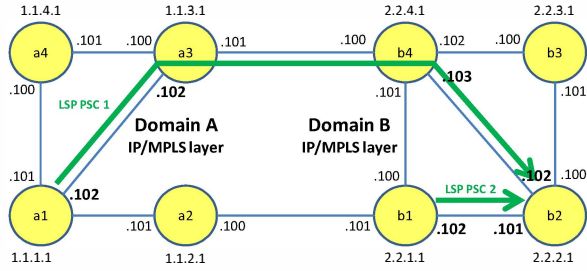

Fig. 5. IP/MPLS layer of the test network, the represented virtual topology considers all the lightpaths represented in Fig. 4, and a lightpath established between each pair of adjacent optical nodes.

Domain A - Optical layer		
LSP LSC id	Node	Card
1	10.1.0.1	10
1	10.1.0.1	1
1	10.1.0.4	1
1	10.1.0.4	2
1	10.1.0.3	1
1	10.1.0.3	10
2	10.1.0.3	11
2	10.1.0.3	2
2	10.2.0.4	1
2	10.2.0.4	11

(a)

Domain A - IP/MPLS layer		
LSP PSC id	Address	Used LSP LSC
1	1.1.1.1	
1	1.1.1.102	1
1	1.1.3.102	1
1	1.1.3.101	2
1	2.2.4.100	2

(b)

Domain B - Optical layer		
LSP LSC id	Node	Card
4	10.2.0.1	10
4	10.2.0.1	2
4	10.2.0.4	2
4	10.2.0.4	3
4	10.2.0.3	1
4	10.2.0.3	2
4	10.2.0.2	1
4	10.2.0.2	10
3	10.2.0.4	10
3	10.2.0.4	3
3	10.2.0.3	1
3	10.2.0.3	2
3	10.2.0.2	1
3	10.2.0.2	11

(c)

Domain B - IP/MPLS layer		
LSP PSC id	Address	Used LSP LSC
1	2.2.4.103	3
1	2.2.2.102	3
1	2.2.2.1	
2	2.2.1.1	
2	2.2.1.102	4
2	2.2.2.101	4
2	2.2.2.1	

(d)

Fig. 6. Database tables: optical layer table at domain A (a); IP/MPLS layer table at domain A (b); optical layer table at domain B (c); IP/MPLS layer table at domain B (d).

IP/MPLS tables are used to associate the established lightpaths with IP/MPLS layer LSPs.

In case of failure of the physical link between ROADM $B3$ and $B2$ (i.e., interface 2 on ROADM $B3$), both IP/MPLS LSPs represented in Fig. 5, are disrupted. However, considering dynamic recovery at the IP/MPLS layer, the information regarding the failure of LSP LSC 1 should be communicated by domain B to domain A thus involving PCEP communication among the two domains. Conversely, using the proposed approach the information regarding the status of the disrupted interface is automatically synchronized between the two databases. After synchronization the controller of domain A can autonomously retrieve the IDs of the IP/MPLS LSPs disrupted by the failure using a set of queries to the databases.

Specifically, a first query is addressed to the optical layer table of domain B, Fig. 6(c), to understand which lightpaths are affected by the failure of the card 2 on ROADM $B3$ (i.e., 10.2.0.3 in Fig. 4). The result of this first query is that LSC LSPs with IDs 4 and 3 are disrupted. The second query uses the output of the first query and is addressed to IP/MPLS layer table of domain B, Fig. 6(d), to understand which PSC LSPs are affected by the failure. The result of this query is that PSC LSPs with IDs 1 and 2 are disrupted. However, PSC LSP 2 is internal at domain B therefore controller of domain A can

discard this information. Conversely, PSC LSP 1 is originated in domain A, therefore the controller of domain A can start the operations to recover this LSP. Indeed, controller of domain A has also access to the intra-domain topology of domain B through the shared database.

The scalability of the proposed procedures has been evaluated in terms of time required to perform the described sequence of queries in a mesh network composed of 100 nodes where 100k PSC LSPs have been emulated and inserted in the IP/MPLS table (reaching around 1 million entries). No scalability issues have been experienced: the time required to perform the last query for lightpath to LSP correlation always remained below 2 ms.

V. Experimental demonstration

The proposed DB-based solution for the exchange of TE information between interconnected TE networks is implemented considering a multi-domain network composed of three domains, each provided with a network controller. Each domain controller implementation, derived from [5], includes a DB to store and maintain TE parameters. Thus, each DB has to store TE info belonging to the local domain as well as to the remote domains.

Two MariaDB 10.1 implementations of type Network Database (NDB) are adopted in this study, both relying on standard SQL-based message exchange of YANG-defined TE parameters. The first DB implementation is a multi-master
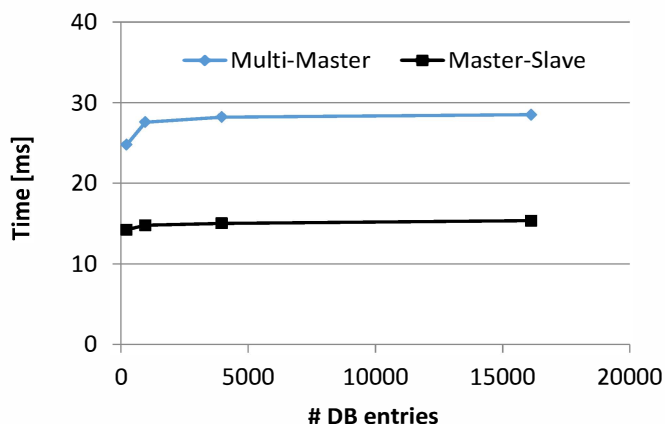
Fig. 7. Average time to exchange (i.e., replicate) one DB entry as a function of the DB size for multi-master and master-slave DB implementations.

(cluster) configuration reproducing a three-domain network scenario. DB_1 and DB_2 run over two Xeon 4-core@3.4GHz, 4G RAM; DB3 over a Xeon 12-core@3.4GHz, 32G RAM. The second DB implementation is a master-slave configuration used to replicate master TE info of a domain to a subscriber remote DB belonging to a different domain. In this implementation, master and slave DB run over two virtualized systems, each provided with 2 cores@2.4GHz and 2G RAM.

In the multi-master implementation, a 3-domain network is considered. In domain 1, a mesh network topology of type $N \times N$ is emulated within DB_1. Four values of N are applied: 8, 16, 32 and 64. In the worst case of $N = 64$, a topology of more than 16k unidirectional TE links is obtained, corresponding to an equivalent number of link entries at DB_1. All DB_1 entries are replicated in DB_2 and DB_3 for end-to-end path computations performed by controllers 2 and 3, respectively. For each value of N, two experiments of update and subsequent replica operations are implemented. 100 repetitions per experiment are performed. Two experiments are considered in this multi-master implementation. In the first experiment, a variable number of entries is modified in DB_1 by controller 1 and replicated to DB_2 and DB_3. The modification is applied to either an unreserved link bandwidth value or to the whole set of link information (no relevant performance differences have been measured in these two cases). In the second experiment, only replication performance is evaluated, considering the scenario where one entire empty DB is added and synchronized. Time performance is measured from the time the first event occurs at the master (e.g., info update) to the time the data is available and consistent on the remote DB (either masters or slaves).

Fig. 7 shows the DB performance in the first experiment. The overall per-entry update and replica procedure over both remote DB_2 and DB_3 is completed in less than 30ms. In this measurement, such value is practically imposed by the slowest synchronized DB. As shown in the figure, no scalability issues occur even if an entire DB of 16k entries is exchanged, providing very stable and predictable time values per entry,

not affected by the DB size. In the second experiment, no relevant differences have been measured. Indeed, the local data update does not significantly impact the overall required time, dominated by the synchronization procedure.

Fig. 7 also shows the experimental results related to the second considered master-slave implementation. Although in this case the DBs are implemented over virtualized systems, better performance is experienced in terms of overall time required to complete the exchange of TE information. Moreover, also in this case, extremely good scalability performance is achieved, showing almost constant time values as a function of the DB size.

VI. CONCLUSIONS

In this work, we propose to exchange TE information among network domains through standard communication among database systems. This represents an attractive, highly scalable and already available alternative to currently adopted communication protocols for exporting TE information (e.g., BGP-LS). For inter-operability purposes, the proposed DB-based solution strictly requires standard YANG data modelling.

Both multi-master and master-slave DB configurations are implemented and evaluated over an emulated large scale network including more than 16k links, showing remarkable scalability performance.

REFERENCES

[1] A. Farrel, J. Drake, and other, "Problem statement and architecture for information exchange between interconnected traffic engineered networks," IETF draft-ietf-teas-interconnected-te-info-exchange-07, May 2016.
[2] A. Giorgetti, F. Paolucci, F. Cugini, and P. Castoldi, "Impact of intra-domain information in GMPLS-based WSONs with hierarchical PCE," in *Tech. Dig. OFC/NFOEC*, Mar. 2012.
[3] H. Gredler, J. Medved, S. Previdi, A. Farrel, and S. Ray, "North-bound distribution of link-state and traffic engineering (te) information using bgp," IETF RFC 7752, Mar. 2016.
[4] A. Giorgetti, "Proactive H-PCE architecture with BGP-LS update for multidomain elastic optical networks [invited]," *J. Opt. Commun. Netw.*, vol. 7, no. 11, pp. 1–9, Nov. 2015.
[5] F. Paolucci, F. Cugini, A. Giorgetti, N. Sambo, and P. Castoldi, "A survey on the path computation element (PCE) architecture," *Communications Surveys Tutorials, IEEE*, vol. 15, no. 4, pp. 1819–1841, Apr. 2013.
[6] F. Paolucci, F. Cugini, G. Cecchetti, and P. Castoldi, "Open database for interconnected traffic engineered multi-layer networks," in *Tech. Dig. OFC*, 2016, p. Th4G.5.
[7] M. Tornatore, F. De Grandi, R. Muoz, R. Martinez, R. Casellas, and A. Pattavina, "Effects of outdated control information in control-plane-enabled optical networks with path protection," *J. Opt. Commun. Netw.*, vol. 1, no. 2, pp. A194–A204, Jul. 2009.
[8] A. Giorgetti, F. Cugini, N. Sambo, F. Paolucci, N. Andriolli, and P. Castoldi, "Path state-based update of PCE traffic engineering database in wavelength switched optical networks," *IEEE Commun. Lett.*, vol. 14, no. 6, pp. 575–577, Jun. 2010.
[9] A. Giorgetti, F. Paolucci, F. Cugini, and P. Castoldi, "Hierarchical PCE in GMPLS-based multi-domain wavelength switched optical networks," in *Tech. Dig. OFC/NFOEC*, Mar. 2011.
[10] J. Vasseur, "A backward-recursive PCE-based computation (BRPC) procedure to compute shortest constrained inter-domain traffic engineering label switched paths," IETF RFC 5441, Apr. 2009.

[11] F. Paolucci, F. Cugini, L. Valcarenghi, and P. Castoldi, "Enhancing backward recursive PCE-based computation (BRPC) for inter-domain protected LSP provisioning," in *Tech. Dig. OFC/NFOEC*, Feb. 2008.

[12] D. King and A. Farrel, "The application of the path computation element architecture to the determination of a sequence of domains in MPLS and GMPLS," IETF RFC 6805, Nov. 2012.

[13] D. Siracusa, S. Grita, G. Maier, A. Pattavina, F. Paolucci, F. Cugini, and P. Castoldi, "Domain sequence protocol (DSP) for PCE-based multi-domain traffic engineering," *J. Opt. Commun. Netw.*, vol. 4, no. 11, pp. 876–884, Nov. 2012.

[14] G. Hernandez-Sola, J. Perello, F. Agraz, S. Spadaro, J. Comellas, and G. Junyent, "Scalable hybrid path computation procedure for PCE-based multi-domain WSON networks," in *Proc. ICTON*, Jun. 2011.

[15] S. Shang, X. Zheng, H. Zhang, N. Hua, and H. Zhang, "A hierarchical path computation element (PCE)-based routing algorithm in multi-domain WDM networks," in *Proc. ACP*, Dec. 2010.

[16] H. Gredler, J. Medved, S. Previdi, A. Farrel, and S. Ray, "North-bound distribution of link-state and TE information using BGP," IETF draft, Jul. 2015.

[17] M. Cuaresma, F. Munoz del Nuevo, S. Martinez, A. Mayoral, O. Gonzalez de Dios, V. Lopez, and J. Fernandez-Palacios, "Experimental demonstration of H-PCE with BPG-LS in elastic optical networks," in *Proc. ECOC*, Sep. 2013.

[18] O. Gonzalez de Dios, R. Casellas, R. Morro, F. Paolucci, V. Lopez, R. Martinez, R. Munoz, R. Vilalta, and P. Castoldi, "First multi-partner demonstration of BGP-LS enabled inter-domain EON control with H-PCE," in *Tech. Dig. OFC*, Mar. 2015.

[19] X. Li, S. Huang, Y. Zhao, M. Zhang, D. Wang, Z. Wang, J. Zhang, and W. Gu, "Combination of distributed and centralized control architecture based on PCE protocol for large scale multi-domain elastic optical networks," in *Proc. ICOCN*, Nov. 2014.

[20] A. Clemm, J. Medved, R. Varga, and other, "A data model for network topologies," IETF draft-ietf-i2rs-yang-network-topo-04, Jul. 2016.

[21] D. Yeung, Y. Qu, and other, "Yang data model for OSPF protocol," IETF draft-ietf-ospf-yang-05, Jul. 2016.

[22] M. Dallaglio, N. Sambo, J. Akhtar, F. Cugini, and P. Castoldi, "YANG model and NETCONF protocol for control and management of elastic optical networks," in *Tech. Dig. OFC*, 2016, p. Th4G.5.