Experimental validation of domain-level gradient-based and hierarchical PCE-based routing and control in heterogeneous optical networks

Rui Lu (鲁 睿)[1*], Xiaoping Zheng (郑小平)[1], Nan Hua (华 楠)[1], Wangyang Liu (刘汪洋)[1], and Xiaohui Chen (陈晓辉)[2]

[1]*Department of Electronic Engineering, Tsinghua University, Beijing 100084, China*

[2]*Fiberhome Telecommunication Technologies Co. Ltd., Wuhan 430074, China*

[*]*Corresponding author: lu-r09@mail.tsinghua.edu.cn*

Received March 21, 2013; accepted June 28, 2013; posted online August 22, 2013

A domain-level gradient-based routing (DLR) algorithm for heterogeneous optical networks with synchronous digital hierarchy and optical transport network domains is proposed and experimentally validated. This algorithm classifies domains into groups with incremental levels on the basis of domain-level partitioning, and guides paths level by level along a gradient on the basis of interdomain routing tree evolution. The proposed algorithm is implemented in the hierarchical path computation element-based control architecture for connection provisioning. Testbeds with commercial and emulated nodes are established to verify the feasibility and performance of the algorithm. Experimental and emulation results show that DLR effectively performs in terms of network blocking probability, real time characteristics, and scalability.

OCIS codes: 060.4250, 060.4251.

doi: 10.3788/COL201311.090601.

The interconnection and intercommunication of heterogeneous optical networks is becoming an important development trend in the telecommunications industry[1,2]. The interdomain routing based on a path computation element (PCE) was proposed by the Internet Engineering Task Force (IETF), and it outperforms other mechanisms because of its high efficiency in interdomain path computation and considerable flexibility in system integration[3−8]. Backward recursive path computation (BRPC) is a regular procedure for interdomain routing processes, calculating optimal paths in a determined sequence of domains[9,10]. Hierarchical PCE architecture is then brought forward to improve performance given that this architecture enhances domain sequence determination[11−13]. Nevertheless, the performance of BRPC remains severely affected by the selection algorithm of domain sequence, which is based on abstract interdomain topology and traffic engineering (TE) information[14]. To solve this problem, researchers have proposed multiple mechanisms[13−15]. For example, scholars developed a lightweight hierarchical PCE-based path computation procedure, which is more suitable for networks that lack interdomain resources[15]. An exhaustive segment path computation scheme was also proposed, in which the parent PCE (pPCE) queries several child PCEs about all possible segments, with consideration for a set of candidate domain sequences; this approach may introduce message overheads between the parent and child PCEs[16]. In Ref. [17], a k-random-paths algorithm that randomly selects border node sequence was proposed; the selection depends on the frequent synchronization of virtual intra-domain link (intra-link) TE information. In the present study, we propose a domain-level gradient-based routing (DLR) algorithm. Rather than determining domain sequence entirely on the basis of abstract TE information, the DLR algorithm selects a cluster of domains with the highest probability of crossing the optimal path and then determines the domain sequence on the basis of the growth and arbitration of different branches during path computation. This algorithm attempts to determine more optimal paths while maintaining as little message overhead as possible. The effectiveness of the algorithm embedded in a hierarchical PCE-based routing and control architecture is verified on testbeds with synchronous digital hierarchy (SDH) and optical transport network (OTN) domains.

The basic idea of the DLR algorithm is to classify domains into different groups, with increasing levels; it is designed to lead a path along the gradient of domain levels, just like "waterfalls". Domain-level gradient-based routing involves three steps: domain-level partitioning, domain set determination, and interdomain routing tree (IDRT) evolution. Firstly, the domain level is partitioned in accordance with the distance of a domain relative to a reference domain. In this letter, this distance is defined as the fewest number of traversed domains between the reference and target domains. As shown in Fig. 1, the source and destination domains are taken as reference domains. Therefore, each domain in a network is labeled as levels $L_{\rm src}$ and $L_{\rm dst}$. Secondly, the domain set for path computation is generated in accordance with two rules. For each domain D_i (i is the number of domains in the entire network and $i \leqslant N_{\rm d}$), if $L_{\rm sum}(D_i) = L_{\rm src}(D_i) + L_{\rm dst}(D_i) \leqslant T_1$, D_i is classified under a temporary domain set $\mathrm{S}_{\rm sel-1}$. Threshold T_1 can be calculated using Eq. (1), with the real number coefficient being $0 \leqslant \eta \leqslant 1$.

$$T_1 = \min\left[L_{\rm sum}(D_i)\right] + \eta \cdot \left\{\max\left[L_{\rm sum}(D_i)\right] - \min\left[L_{\rm sum}(D_i)\right]\right\}. \quad (1)$$

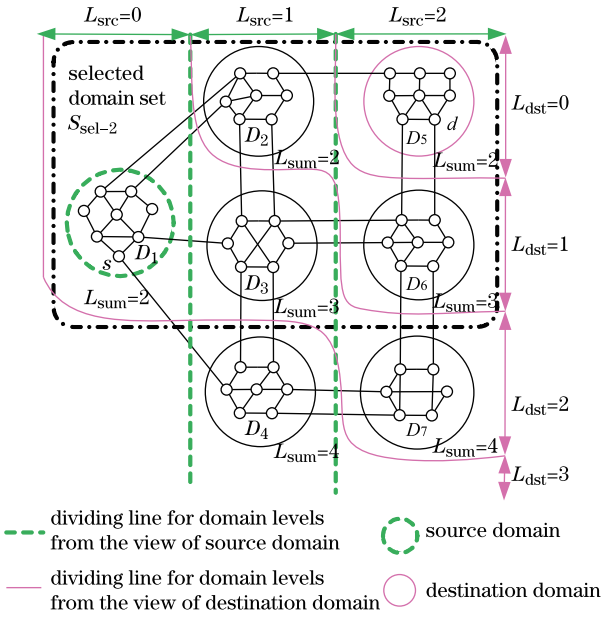

Fig. 1. (Color online) Domain level partition and domain set selection.

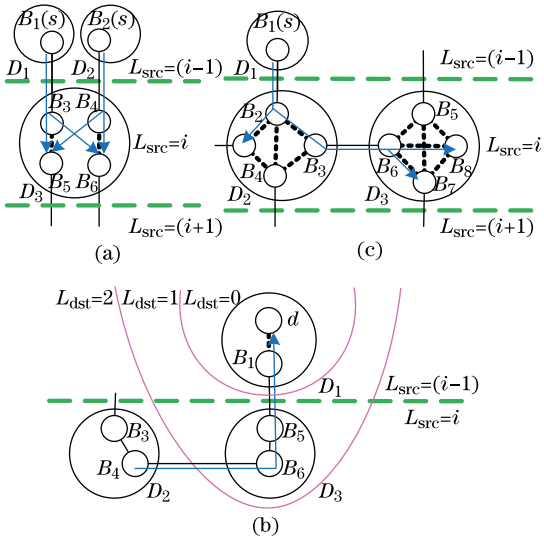

Fig. 2. (Color online) Three types of IDRT growth: (a) Type-I; (b) Type-II; (c) Type-III.

For each domain D_i (i is the number of domains in S_{sel-1} and $i \leqslant N_{sel-1}$) in S_{sel-1}, if all the neighbor domains of D_i (D_i is neither the source nor the destination domain) are in the same level that differs from the level of D_i, then it is removed from S_{sel-1}, and the remaining domains constitute domain set S_{sel-2}. In the example in Fig. 1, $\eta = 0.5$, Threshold$_1$ = 3, and S_{sel-2}= {D_1, D_2, D_3, D_5, D_6}. The succeeding path computation is implemented in the scope of domains in S_{sel-2} and all the interdomain links (interlinks) that connect them. Thirdly, IDRT evolution entails two basic operations: IDRT growth and pruning. Figure 2 shows that the branches of IDRT expand from the source node (s), go across intermediate border nodes (B_k), and arrive at the destination node (d). As indicated by the priority of subfigures from Figs. 2(a) to (c), IDRT growth is classified into three types: (a) Type-I, in which IDRT grows in the ascent direction of L_{src}; (b) Type-II, wherein if $L_{src}(D_i) \geqslant L_{src}(D_d)$ (D_d is the destination domain), IDRT grows in the descent direction of L_{dst}; and (c) Type-III, in which IDRT traverses domains at the same level. Type-III growths are implemented after each step of Type-I or Type-II growth, which is determined on the basis of L_{src} or L_{dst}, respectively. Each branch of IDRT goes across a domain only once to avoid the formation of loops on a path.

As shown in Fig. 3, IDRT pruning is implemented when different branches converge at a node or an interlink. A short branch with low cost is retained, whereas the others are excluded from IDRT. As indicated in the priority of subfigures from Figs. 3(a) to (e), IDRT pruning is categorized into five types: (a) Type-I, in which pruning is conducted for branches after interlevel growth and arrival at the same border node; (b) Type-II, wherein pruning is implemented for branches after interlevel or intra-level growth and arrival at an interlink that connects border nodes at the same level; (c) Type-III, where pruning is executed for branches with at least one of them having experienced interlevel growth within a short period and arriving at a border node; (d) Type-IV, in which pruning is implemented for branches arriving at a destination node; (e) Type-V, in which pruning is executed for branches that cannot reach the destination domain after all types of growth.

An example of IDRT evolution is presented in Fig. 4. This process is implemented in a subnetwork determined by S_{sel-2} in Fig. 1. In this example, the number of hops is considered as the cost and length of a branch. As shown in Fig. 4(a), Type-I growth is firstly carried out from the source node to the border nodes in D_1. Secondly, as shown in Fig. 4(b), Type-I and Type-III growth is implemented from L_{src}= 0 to L_{src}= 1, and Type-I pruning is implemented to guarantee that the short branches arriving at the border nodes of D_2 and D_3 are retained. Given that the branches produced after Type-III growth are longer than those generated after

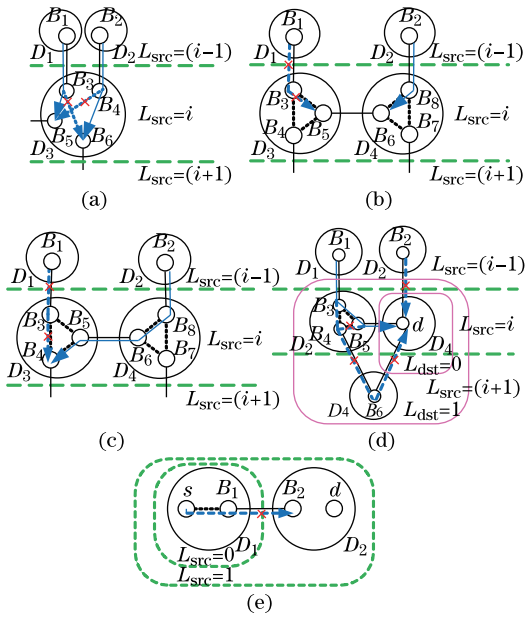

Fig. 3. (Color online) Five types of IDRT pruning: (a) Type-I; (b) Type-II; (c) Type-III; (d) Type-IV; (e) Type-V.

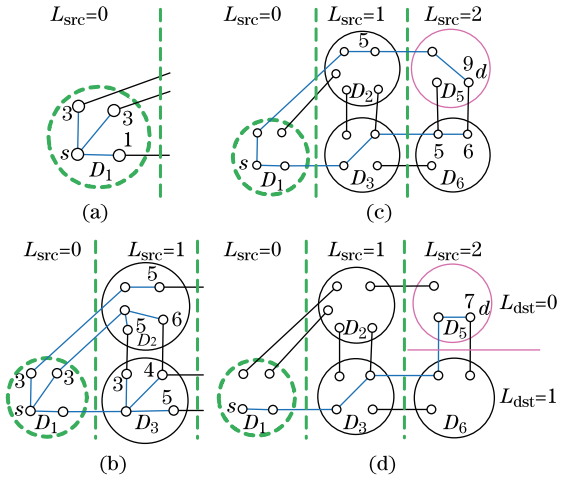

Fig. 4. (Color online) Example of IDRT evolution.

Type-I growth, Type-II and Type-III pruning is executed and no branch is expanded between D_2 and D_3. Thirdly, as presented in Fig. 4(c), Type-I growth and Type-I pruning from D_2 to D_5 and from D_3 to D_6, respectively, are executed. Because D_5 is the destination domain, the branch that traverses $D_1 - D_2 - D_5$ arrives at destination node d and forms a path with a length of 9. Finally, as presented in Fig. 4(d), Type-II growth from D_6 to D_5 is executed to reach the destination node and form a path with a length of 7. After comparison by Type-IV pruning, the branch that traverses the domain sequence $D_1 - D_3 - D_6 - D_5$ is retained because its cost is lower than that incurred by the branch that traverses the domain sequence $D_1 - D_2 - D_5$. Type-V pruning is not implemented because all branches can arrive at destination node d.

The DLR algorithm can be accomplished under hierarchical PCE-based interdomain management and control architecture for connection provisioning[18]. As illustrated in Fig. 5, this architecture comprises four entities: traditional generalized multi-protocol label switching (GMPLS) control planes (CPs)[19], hierarchical PCEs (H-PCEs), hierarchical interdomain connection control element (H-ICCEs), and a network management system (NMS). Traditional CPs in heterogeneous optical networks are usually independent from one another. Each CP is in charge of provisioning through an intra-domain label switch path (LSP) via resource reservation protocol (RSVP) signaling and intra-domain traffic engineering database (TED) management. H-PCEs and H-ICCEs construct a uniform interdomain control layer over separated CPs to accomplish interdomain routing and connection provisioning. The pPCE is responsible for interdomain path computation, such as domain-level partitioning, domain set determination, and IDRT evolution. The cPCE is in charge of path segment computation within a domain, between source, as well as the destination and border nodes based on the intra-domain TED synchronized from CPs in the corresponding domain. The pPCE interacts with the cPCEs in a request/response signaling process based on the extended path computation element communication protocol (PCEP)[20,21]. The pPCE sends path segment computation requests to cPCEs in accordance with IDRT evolution rules. The cPCEs calculate the shortest path between two end nodes of a segment and send the costs to the pPCE. Two-layer H-ICCEs are responsible for interdomain connection provisioning, protection, and restoration. Similar to H-PCEs, the ICCE on the upper layer is denoted as the parent ICCE (pICCE) and the ICCE within each domain on the lower layer is denoted as the child ICCE (cICCE). The connection setup is a stitching process of multiple intra-domain segments. The pICCE takes charge of configuring the interdomain tributary interface card, thereby guaranteeing that the interfaces on both ends of the interlink communicate at the same wavelength and time slot. The cICCE is used to accomplish the intra-domain LSP provisioning. The cICCE communicates with the corresponding CP and configures the intra-domain aggregate interface cards and cross-connection cards on the basis of the path computation results within this domain. NMS is employed to launch end-to-end connection provisioning and to manage the network. As presented in Fig. 5, the interactions among these entities for an interdomain connection setup are of six types: (1) connection requests and responses between NMS and pICCE; (2) interdomain path computation requests and responses between the pICCE and pPCE; (3) intra-domain path computation interactions between the pPCE and cPCEs; (4) intra-domain LSP setup interactions between the pICCE and cICCEs; (5) intra-domain communications between the cICCEs and corresponding cPCEs to acquire LSP information; and (6) intra-domain interactions between the cICCEs and corresponding CPs to configure the tributary and aggregate interface cards within the domain. For a multi-layer domain, both the light-path and upper layer of the electronic path should be established.

Three testbeds are used to verify the effectiveness of the DLR algorithm and the H-PCE-based routing and control architecture. As shown in Fig. 6(a), all the equipment of management, control, and data planes are connected to a virtual local area network. For each domain, a embedded interdomain control element (ICE) is developed and operated[22]. This element integrates the cPCE and cICCE in the corresponding domain and

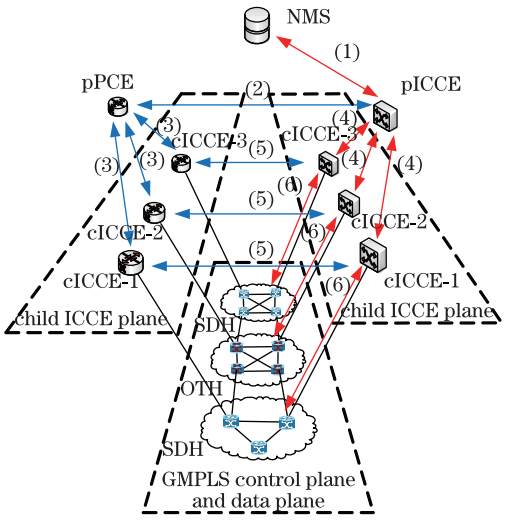

Fig. 5. (Color online) Control architecture for connection provisioning.

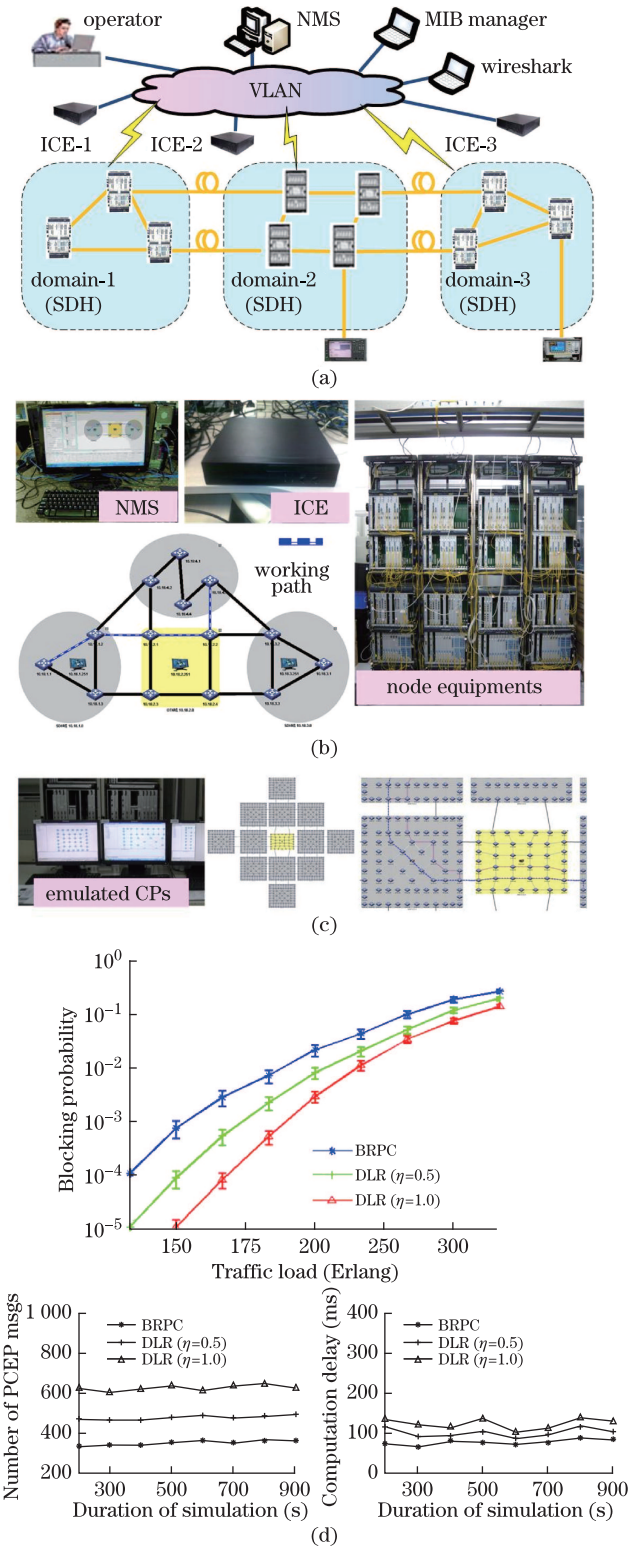

Fig. 6. (Color online) DLR testbed for heterogeneous optical networks.

creates a backup of the pPCE and pICCE. For each connection request, the higher layer PCE and ICCE in the ICE of the source domain is activated for interdomain routing and connection control. The changes in interdomain TE information are broadcast to other ICEs after the setup or release of connections. NMS is run on a server for overall network management. Wireshark® software is utilized as a PCEP protocol analyzer. In addition, the management information base manager is employed to inquire and monitor the TED information of CP and the subnet management system. SDH and OTN test instruments are connected to the data plane of the network to monitor the status of light-paths on the physical layer.

Table 1. Evaluation of Real-time Characteristics

Algorithm	DLR	BRPC
Path Computation (ms)	< 150	< 100
Connection Provisioning (ms)	$< 700*$	$< 560*$

*The delay of automatic discovery of interlayer resources is disregarded.

As presented in Fig. 6(b), the first testbed is established on a heterogeneous optical network by using 10 pieces of FonsWeaver® 780 SDH equipment and 4 pieces of FONST3000® OTN equipment. This network is divided into 4 domains. For the SDH domain, the add/drop interfaces for the connection are in a granularity of VC-4 (155 Mb/s). The STM-64 frames or VC-4-64c are implemented on the intra-links. For the OTN domain, the add/drop interfaces for the connection are in granularities of ODU-1 (2.5 Gb/s), ODU-2 (10 Gb/s), and wavelength (10 Gb/s). Each intra-link can carry 4 wavelengths. Each interlink carries STM-16 SDH frames. As shown on the lower left corner of Fig. 6(b), a path from the western SDH domain to the northern SDH domain is set up. Instead of having the interlinks pass through the SDH domain, as indicated by the algorithm of the fewest number of traversed domains (typically used in BRPC), the path initially traverses the OTN domain and then turns to the northern SDH domain. Thus, the DLR algorithm determines a shorter path with a length of 4 in this domain sequence. The real-time characteristics of DLR and hierarchical PCE-based control are evaluated (Table 1). Given that the DLR algorithm calculates the path in more domains than does the BRPC algorithm, the delay in path computation and connection setup of the DLR algorithm is slightly longer.

The second testbed is established on the basis of a multi-threading emulated platform with 1 000 ASON nodes, each emulated by a thread. The scalability of DLR and H-PCE-based control architecture is verified on this platform. The network comprises 13 domains, one of which is an OTN domain and the rest are SDH domains. Each emulated domain is implemented on a server. The network resources and the capacity of each node and link are the same as those on the first testbed. As shown in Fig. 6(c), connections traversing 5 domains are set up within the restriction of signal delay and overheads.

The blocking probability of DLR is compared with that of traditional BRPC on a simulated testbed with the same topology shown in Fig. 1 (Fig. 6(d)). A total of 7 domains, 49 nodes, and 87 links exist in the network for simulation. Each intra-link accommodates 32 wavelengths and each interlink accommodates 64 wavelengths. The connection requests arrive at the network following a Poisson distribution and are uniformly distributed among each node pair. The duration of each request is exponentially distributed with the same mean time. The traffic load is modified by changing the average holding time of connections, and the average interarrival time is set as

1 s. The border nodes of each domain are capable of full wavelength conversion. The simulation results show that the DLR algorithm achieves a significantly lower blocking probability than does the BRPC algorithm. Given its capability to calculate a shorter path while traversing a greater number of domains compared with BRPC, the DLR algorithm is particularly suitable for connection provisioning in multi-domain network with abundant interdomain links and bandwidth resources. In addition, the blocking probability resulting from different η is presented in Fig. 6(d). The larger the η, the greater the number of domains included in domain set $S_{\text{sel}-2}$, and the lower the blocking probability achieved by the DLR algorithm.

We also analyze path computation delays, and that of the DLR algorithm is subject to the number of levels (N_{level}) for the domains in $S_{\text{sel}-2}$. Path computations for domains at the same level are implemented in parallel. The round trip time (RTT) between the parent and child PCE is within 13 ms. The total delay for a path computation can be approximated as ($N_{\text{level}} \times$ RTT). Although the number of PCEP messages of DLR is greater than that of BRPC, the delay of DLR is only slightly larger than that of BRPC (Fig. 6(d)). Moreover, race conditions are observed during experimentation. This phenomenon is caused by the interdomain TED synchronization delays between the ICE and parent PCE and other ICEs. The maximum delay for the two previous testbeds is within 50 ms, which may result in connection failures of 7% to 18%.

In conclusion, a domain-level gradient-based routing algorithm and hierarchical PCE-based routing and control architecture are proposed for interdomain connection provisioning in heterogeneous optical networks. Experimental results derived from testbeds with commercial and emulated nodes show that the DLR algorithm and hierarchical PCE-based control architecture satisfy the functionality requirements of connection scheduling in heterogeneous optical networks with good real-time characteristics and scalability. Moreover, the DLR algorithm and hierarchical PCE-based control architecture can achieve lower blocking probability than can traditional BRPC algorithm and control architecture despite an acceptable increase in signal delay and overhead.

This work was supported by the "863" Project of China (Nos. 2012AA011301 and 2009AA01Z254), the National "973" Program of China (Nos. 2010CB328203 and 2010CB328205), and the National Natural Science Foundation of China (No. 61201188).

References

1. M. Chamania and A. Jukan, IEEE Commun. Surv. Tut. **11,** 33 (2009).
2. L. Zhou and C. V. Saradhi, in *Proceedings of 10th IEEE Singapore International Conference on Communication Systems* 1 (2006).
3. W. Sun, P. Li, C. Li, and W. Hu, Chin. Opt. Lett. **11,** 010601 (2013).
4. A. Farrel, J. P. Vasseur, and J. Ash, "A Path Computation Element (PCE)-based architecture," RFC 4655 (2006).
5. T. Takeda, R. Sugiyama, E. Oki, I. Inoue, S. Kohei, K. Shindome, K. Fujihara, and S.-I. Kato, in *Proceedings of 34th European Conference on Optical Communication* 1 (2008).
6. R. Lu, L. Wang, Q. Li, X. Wan, C. Yang, N. Hua, Q. Jin, S. Shang, X. Zheng, H. Zhang, Y. Guo, X. Chen, and L. Liao, in *Proceedings of Optical Fiber Communication Conference and Exposition and the National Fiber Optic Engineers Conference (OFC/NFOEC)* 1 (2011).
7. S. Dasgupata, J. C. de Oliveira, and J.-P. Vasseur, IEEE Network **21,** 38 (2007).
8. A. P. Bianzino, J. Rougier, S. Secci, R. Casellas, R. Martinez, R. Munoz, N. B. Djarallah, R. Douville, and H. Pouyllau, in *Proceedings of TridentCom 2009. 5th International Conference on Testbeds and Research Infrastructures for the Development of Networks & Communities and Workshops* 1 (2009).
9. I. Nishioka, Y. Iizawa, and S. Araki, Proc. SPIE **6784,** 67840T (2007).
10. Y. Lu and L. Hou, Chin. Opt. Lett. **10,** 040602 (2012).
11. J. P. Vasseur, R. Zhang, N. Bitar, and J. L. Le Roux, "A backward-recursive PCE-based computation (BRPC) procedure to compute shortest constrained inter-domain traffic engineering label switched paths," RFC 5441 (2009).
12. F. Paolucci, F. Cugini, L. Valcarenghi, and P. Castoldi, in *Proceedings of Optical Fiber Communication Conference* OTuA5 (2008).
13. H. Matsuura, N. Morita, T. Murakami, and K. Takami, in *Proceedings of Global Telecommunications Conference* 2072 (2005).
14. 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 (2012).
15. A. Giorgetti, F. Paolucci, F. Cugini, and P. Castoldi, in *Proceedings of National Fiber Optic Engineers Conference* NTuC4 (2011).
16. A. Giorgetti, F. Paolucci, F. Cugini, and P. Castoldi, in *Proceedings of National Fiber Optic Engineers Conference* NTu2J.2 (2012).
17. S. Shang, N. Hua, L. Wang, R. Lu, X. Zheng, and H. Zhang, Optical Switching and Networking **8,** 235 (2011).
18. R. Lu, X. Zheng, and N. Hua, Journal of Tsinghua University (in Chinese) (to be published) (2013).
19. Z. Du, Y. Lu, and Y. Ji, Chin. Opt. Lett. **10,** 020604 (2012).
20. J. P. Vasseur and J. Le Roux, "Path computation element (PCE) communicatioin protocol (PCEP)," RFC 5440 (2009).
21. J. Ash and J. L. Le Roux, "Path computation element (PCE) communication protocol generic requirements," RFC 4657 (2006).
22. R. Lu, X. Zheng, N. Hua, Q. Jin, W. Liu, and X. Chen, Optical Communication Technology (in Chinese) **6,** 1 (2013).