Optical Switching and Networking 11 (2014) 119–135

Contents lists available at ScienceDirect

Optical Switching and Networking

journal homepage: www.elsevier.com/locate/osn

Multi-domain optical routing: Is there life beyond extending BGP?

Eva Marín-Tordera, Xavier Masip-Bruin *, Marcelo Yannuzzi, René Serral-Gracià

Advanced Network Architectures Lab (CRAAX), Technical University of Catalonia (UPC), Vilanova i la Geltrú, Spain

ARTICLE INFO

Available online 5 July 2013

Keywords:
Multi-domain routing
Optical networks
Multi-layer networks
Control plane

ABSTRACT

The design of new inter-domain optical routing protocols may start from scratch, or on the contrary exploits all the research already developed in IP networks with the Border Gateway Protocol (BGP). Even though the network premises under which BGP was conceived have drastically changed, the pervasive deployment of BGP makes almost impossible its replacement, hence everything indicates that BGP-based routing will remain present in the coming years. In light of this, the approach often used for distributing reachability information and routing inter-domain connections below the IP layer has been to propose extensions to the BGP protocol, what unfortunately exports all well-known BGP weaknesses to these routing scenarios. In this paper we deeply analyze all these problems in order the reader to get a clear idea of the existing limitations inherent to the BGP, before exploring the routing problem in optical networks. Then, focusing on the optical layer we will demonstrate that current optical extensions of BGP do not meet the particular optical layer constraints. We then propose minor, though effective, changes to a path vector protocol overall offering a promising line of work and a simple solution designed to be deployed on a multi-domain and multi-layer scenario.

1. Introduction and context

The Internet is a decentralized collection of networks, grouped and interconnected in the form of domains or Autonomous Systems (ASs). Each AS typically represents a pool of networks, managed by a single authority, and under a common routing policy. At present, the Internet has approximately 43,000 ASs [5], each of which uses one or more Interior Gateway Protocols (IGPs) for routing within the AS. However, when a node u in a domain AS_U needs to communicate with a node v located in another domain AS_V, the interior routing protocols in AS_U are not sufficient. The process that handles the exchange of routes among ASs is referred to as inter-domain routing. The Border Gateway Protocol (BGP), a path vector protocol, is the "de facto" protocol for inter-domain routing in IP networks. Despite its well-known limitations mainly rooted on the fact that BGP was designed to provide reachability information, BGP is yet the protocol used on the IP network layer to guarantee worldwide connectivity. This wide acceptance of BGP motivated that this protocol was envisioned as the foundation to provide a solution for inter-domain routing at the optical layer. However, some critical aspects are hindering and complicating the adoption of BGP-based protocols in the optical layer. In particular we can mention the following: (i) any extension of BGP over optical networks, such as the Optical Border Gateway Protocol (OBGP) [2], inherits the non-solved problems of BGP; (ii) optical networks do not build paths but rather lightpaths, hence adding a new component, the wavelength, in the routing process that is not considered in the traditional BGP; and (iii) BGP does not consider physical impairments, metric that undoubtedly must be considered when selecting a lightpath. Therefore, any research effort in providing a solution for inter-domain routing in optical networks must be feed on the long record of research existing

* Corresponding author. Tel.: +34 938967280.
E-mail addresses: eva@ac.upc.edu (E. Marín-Tordera), xmasip@ac.upc.edu, xmbruin@gmail.com (X. Masip-Bruin), yannuzzi@ac.upc.edu (M. Yannuzzi), rserral@ac.upc.edu (R. Serral-Gracià).
URL: http://www.craax.upc.edu (X. Masip-Bruin).

1573-4277/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.osn.2013.05.005

for BGP at the IP layer, hence trying to benefit from the lessons learned in the past. To this end, we believe that a solid background on inter-domain routing concepts is radically necessary to position any potential research contribution on multi-domain routing at the optical layer.

The remainder of this paper is organized as follows. Section 2 reviews different aspects of BGP, including its weaknesses not yet solved. Section 3 describes current BGP-based optical extensions and also proposes new optical routing models addressing problems reviewed in Section 2. Section 4 summarizes the effects of physical impairments when computing inter-domain optical light-paths. In Section 5 we review the recent work on control plane extensions for optical inter-domain routing with physical impairments and we present our proposed control plane solution. Then, in Section 6 we briefly introduce the problem of inter-domain routing in the context of a multi-layer network. Performance studies of the new optical routing models proposed in Section 3 are outlined in Section 7. Finally, Section 8 concludes the paper.

2. The legacy concepts

This section aims at familiarizing the reader with usual inter-domain problems and limitations in the context of IP networks, mainly focusing on BGP pros and cons, hence illustrating both the main rules and procedures related to the overall BGP performance and the problems not yet solved at the IP layer, that must be avoided if pushing for a BGP-oriented strategy to address the multi-domain routing problem in optical networks.

2.1. Route dissemination

In BGP, for scalability and confidentiality reasons, the routing information managed and exchanged among ASs is highly condensed. Different from link-state routing protocols, which maintain the topological state of the network, BGP only handles AS-level paths for any possible destination. An AS-level path is composed of a set of attributes, including an ordered sequence of AS numbers (a vector of ASs) that need to be traversed to reach a destination. This routing paradigm is thus called *path vector routing*.

To illustrate the exchange and dissemination of these vectors (called *path vectors*) let us consider the example depicted in Fig. 1. Suppose that AS0 has allocated the IP prefix 10.0.0.0/8, so AS0 advertises the prefix to AS1 and AS2 indicating that it can be reached through the path vector $\overrightarrow{p_0} = [0]$. Both AS1 and AS2 process the routing advertisement received from AS0, prepend their own AS number to the vector of ASs, and advertise the prefix with the corresponding AS-path to each other. More precisely, AS1 advertises to AS2 that the prefix 10.0.0.0/8 can be reached through $\overrightarrow{p_1} = [1\ 0]$, and AS2 does the same and advertises to AS1 the route $\overrightarrow{p_2} = [2\ 0]$. Accordingly, AS1 and AS2 receive two advertisements to reach 10.0.0.0/8; they choose one of them (only one route is advertised in BGP), and in the case of AS2, it advertises its best route to AS3. Suppose that AS2 chooses the route [0] over [1 0], so domains AS3 and AS4 will learn the routes $\overrightarrow{p_3} = [2\ 0]$, and $\overrightarrow{p_4} = [3\ 2\ 0]$, respectively. Fig. 1 shows the state of the routing tables of each AS once the dissemination process for the prefix 10.0.0.0/8 has converged (note that the best route is denoted in bold).

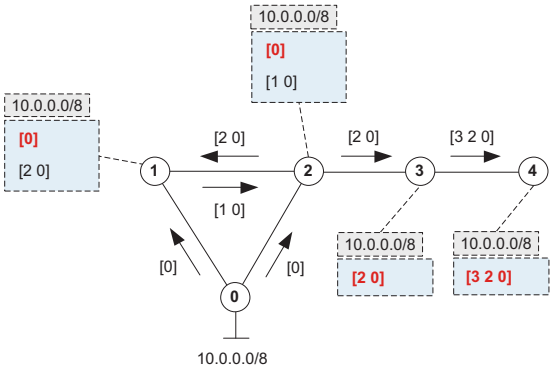

Fig. 1. A simplified version of the exchange and dissemination of information of a path vector routing protocol.

The example above shows that the routing paradigms of distance vector and path vector protocols have similarities. Whereas distance vector protocols choose routes according to the shortest distance to a destination, a path vector protocol will generally choose the route that traverses the least number of ASs. The term "generally" mentioned before is because the AS-path length (a rough sense of distance) is the attribute that is typically considered during the route selection process, but is not the only one. Any AS can change the preference of a route and override the AS hop count, and even change the attributes of the routes it uses and advertises to other devices based on commercial interests and the policies locally configured on each router. The combination of these features allows domains to enforce their routing policies, enabling control over their traffic according to their criteria. These features have made path vector routing one of the major enablers of the expansion experienced by the Internet over the last 15 years. Indeed, its routing model provides sufficient flexibility so that the policies of independent domains can be reflected on the routing system, while preserving the autonomy, confidentiality, and administrative limits of routing domains. With path vector routing, neither the internal topology of a domain nor its interior routing state is disclosed to other domains. This routing model has shown to be scalable year after year, systematically, thanks to the level of aggregation in the information and state maintained in the routing tables. On the other hand, a large list of studies show that BGP suffers from several problems, some of which are due to the implementation decisions while BGP was developed, while others are inherently caused by the utilization of path vectors. In brief, the list includes slow convergence [7], high churn rate of route advertisements [10], limited capabilities to achieve Traffic Engineering (TE) objectives [11], the potentially conflicting nature of BGP routing policies[1] [6], the inability to find and provide paths that guarantee the performance and reliability of inter-domain communications, security vulnerabilities [9], and more.

[1] This is because routing policies are managed independently and without any global coordination among domains.

2.2. Policy-based routing

Previous section has defined the way route information is advertised within a BGP network. However the presented strategy is still missing a critical aspect, motivated by how ASs interact each other. In fact, in order to clearly understand how inter-domain routing information is advertised in the Internet the commercial relationships between ASs must be considered. There are two major types of relationships, namely, *customer–provider* and *peer-peer*. These correspond to the two different traffic exchange agreements between neighboring domains. The former applies when a domain buys Internet connectivity from a provider. The latter, on the other hand, applies when two providers that exchange a significant amount of traffic, agree to connect directly to each other to avoid transiting through, and thus pay, a third-party provider. The peers share the costs of the connection between them, so there is no customer–provider relationship in this case. These two types of commercial relationships impose constraints on the forwarding policy of domains. To illustrate these constraints and the reasons why they are applied in practice, let us consider three domains, AS_i, AS_j, and AS_k, such that AS_i is adjacent to AS_j and the latter is adjacent to AS_k.

- Suppose that AS_j is a customer of AS_i. Then, AS_j will forward the traffic received from AS_i to its customers, but never to its peers or other providers, since AS_j will not be willing to provide transit to the greater Internet to its peers and providers.
- Suppose that AS_j is a provider of AS_i. Then, AS_j will forward the traffic received from AS_i to its customers, providers and peers, since AS_j will provide transit to its customers without restrictions.
- Suppose now that AS_j is a peer of AS_i. Then, AS_j will forward the traffic received from AS_i to its customers but never to its providers or peers, since as in the first case, AS_j will not be willing to pay for the transit to the greater Internet of its peers.

Table 1 summarizes the conditions under which AS_j will forward the traffic received from AS_i to AS_k. These policies are usually referred to as *valley-free routing policies*, and they are enforced by means of route filtering. In this subsection, we describe the use of these filters and the way they are applied to control inter-domain routing advertisements. Indeed, we will show that due to the route filtering, the topologies that can be built by domains from path-vector routing advertisements differ, leading to inconsistencies in the inference of the interconnectivity of the network (cf. Fig. 5).

Table 1
Valley-free routing policies applied by domain AS_j for the transit from domain AS_i to domain AS_k through AS_j ($AS_i \rightarrow AS_j \rightarrow AS_k$).

Commercial relationships	AS_j is a customer of AS_k	AS_j is a provider of AS_k	AS_j is a peer of AS_k
AS_i is a provider of AS_j	×	✓	×
AS_i is a customer of AS_j	✓	✓	✓
AS_i is a peer of AS_j	×	✓	×

2.2.1. Generation and storage of routing state

We proceed now to explain the internals of the route filtering process, and how the routing information is handled, processed, stored, and advertised by a path vector router. To this end, consider the simplified version of a path vector router depicted in Fig. 2, and assume that the routing information flows from left to right— Fig. 2 is an adapted version of a figure introduced by Quoitin et al. in [12]. The left-hand side of Fig. 2 shows that, for each neighbor, the router has configured a set of inbound filters, which are utilized both to select and manipulate the advertisements received from each neighbor. For example, the router might have a policy configured that states that it must never use neighbor X to send traffic to *d*, so the router will filter any route toward *d* from the advertisements received from neighbor X.[2] The advertisements that pass the filtering process are stored in the Routing Information Base (RIB), which is basically the routing table where all candidate routes are maintained. When the RIB contains at least two routes for the same destination, the router needs to choose the best route, which is the one that will be used to forward the traffic toward the destination. This selection is represented as the "Path Computation" process in Fig. 2. The route resulting from this decision is installed in the Forwarding Information Base (FIB), which is the database that holds the routes that will be used by the router to forward the traffic. Whereas a BGP router located in an Internet exchange point may currently have more than 11×10^6 route entries in its RIB, after the path computation process, the FIB will have around 0.33×10^6 entries (see, e.g., the statistics for AS6447 in [5]). Once the router has filtered and selected the routes from the advertisements received, the router will forward traffic from one of its neighbors to another if, and only if, the transit between these two through the router is valley-free. Therefore, the router will advertise upstream a set of "valley-free routes" which must also match its inbound traffic policy. This is enforced by means of the outbound filters shown on the right-hand side of Fig. 2.

2.2.2. Path computation and policy control

Let us focus now on the "Path computation" module in Fig. 2. Since the selection of the best route in a domain is ruled by the economical relationships with its neighbors, the inter-domain routing preferences of a domain are determined as follows. A domain will prefer customer routes over peer routes or provider routes, independent of the AS-path length. The reason is simple; a domain will typically charge its customers for the traffic sent to them, whereas it will be charged by its providers for the traffic sent through them. Moreover, a domain will prefer peer routes over provider routes, because if the net traffic is relatively balanced in a peering link, then none of the peers will be charged by the other. Fig. 3 illustrates the path computation process of a

[2] Notice that since routing advertisements flow from left to right in Fig. 2, the traffic for destinations contained in those advertisements will flow in the opposite direction.

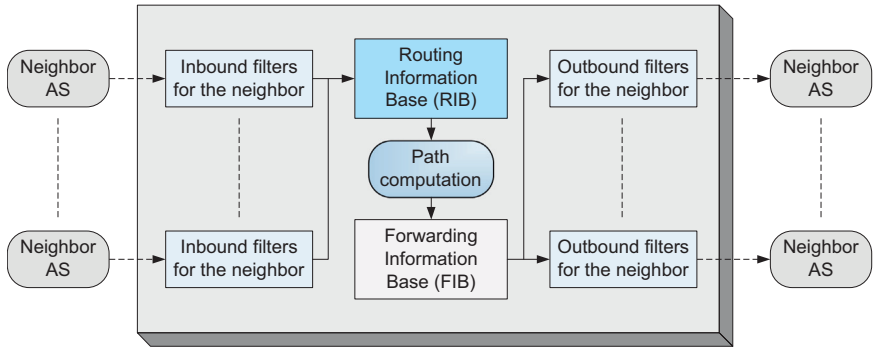

Fig. 2. A simplified view of the information flow and storage within a path vector router (adapted from a figure contained in [12]).

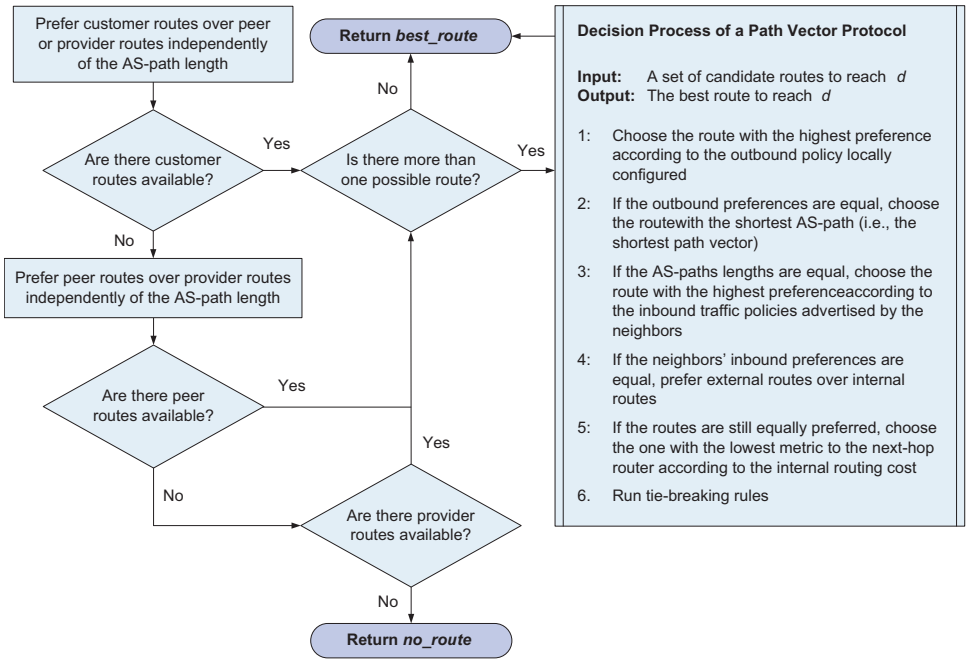

Fig. 3. Path computation process of a path vector routing protocol.

generic path vector routing protocol toward a destination *d*. Since a router will typically have several routes to reach *d*, the routing protocol needs to choose one, i.e., the best path to *d*. The algorithm that a path vector router runs to make this selection is referred to as the *decision process*. The sequence of steps shown on the right-hand side of Fig. 3 summarizes this process. Each step in the process is used to break ties when the routes being compared were equally good in the previous step. The reader who wants to go deeper into the details can consult the decision process adopted in the BGP standard [13].

2.3. Main problems of path vector protocols

After the short description offered in the last subsection on the BGP internals, we propose in this subsection a brief review of the BGP weaknesses that still remain unsolved. Indeed, over the last years, multiple studies have examined the weaknesses of the BGP protocol [13], and exposed its limitations to meet many of the routing requirements for future multi-domain networks. In this subsection, we describe a set of 5 critical issues.

Routing limitations — With the current implementation of BGP, a router has no means to find inter-domain paths subject to constraints, such as paths with a certain amount of available bandwidth, or with bounded delay, bounded losses, or combinations of these. Indeed, the protocol does not handle a "load" component, so the BGP decision process is unable to avoid congested paths. The protocol also lacks multi-path routing capabilities, and therefore, the traffic cannot be balanced among different paths or even it is impossible to select a path with a minimum Quality of Service (QoS). To address these issues, works such as [15] have proposed to integrate QoS into the inter-domain routing system, while others such as [14] have addressed the multi-path problem. Despite the many efforts and over a decade of work, none of the proposals have been made so far in the areas of QoS and multi-path has become widely deployed. Instead, providers have preferred to simplify the operation and maintenance of their networks and have systematically

relied on capacity overprovisioning for improving the performance and reliability of their services.

Slow convergence and churn — Depending on the location of the origin and where the observation is made, a BGP convergence might vary between tens and several hundreds of seconds [7]. This slow convergence is mainly caused by the *path exploration* performed by BGP. An example of this process is shown in Fig. 4. Suppose that AS1 can reach the prefix 10.0.0.0/8 in AS3 through three candidate paths, namely, [2 3], [2 4 3], and [2 5 4 3]. Assuming that each AS chooses the shortest AS-path to reach prefix 10.0.0.0/8, and that AS5 chooses the path through AS4, the following set of events occurs when AS3 loses connectivity with 10.0.0.0/8. AS3 sends withdraw messages both to AS2 and AS4. The withdraw messages reach AS2 and AS4 roughly at the same time, and after processing the message, both AS2 and AS4 explore their best alternative and advertise the changes upstream. In particular, AS2 tries the path [4 3], without knowing that AS4 has received a withdraw for 10.0.0.0/8 too. When AS1 receives the update from AS2, it attempts its next best path, i.e., path ② in Fig. 4. Notice that AS1 will start forwarding traffic toward 10.0.0.0/8 through path ② roughly at the same time that AS2 receives a withdraw for the path [4 3] from AS4. It is easy to see that both AS2 and AS1 will then try the path through AS5 (path ③), until they finally realize that the prefix is unreachable. The problem is that upstream domains have no way to infer that the alternative paths are also affected by the failure. This leads to a time consuming process, which may even take minutes to stabilize in large size networks. The process of exploring paths has other negative effects, such as the amount of messages (churn) that generates [10]. Different initiatives have proposed solutions to limit the path exploration process. Most of them are based on tagging additional information to the withdraws sent by the routing protocol (see, e.g., [4]). However, none of the existing proposals have been sufficiently appealing and easy to integrate in practice so as to become adopted.

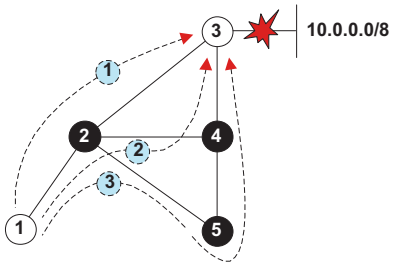

Fig. 4. Path exploration: when the link connecting the network 10.0.0.0/8 fails, upstream domains such as AS1 will explore the alternative paths until they realize that the network is unreachable.

Security vulnerabilities — The BGP protocol lacks both path and origin authentication, and therefore, a BGP router can be perfectly used to advertise any possible (prefix, path vector) pair to the Internet. This makes the inter-domain routing system extremely vulnerable to certain attacks, since both IP

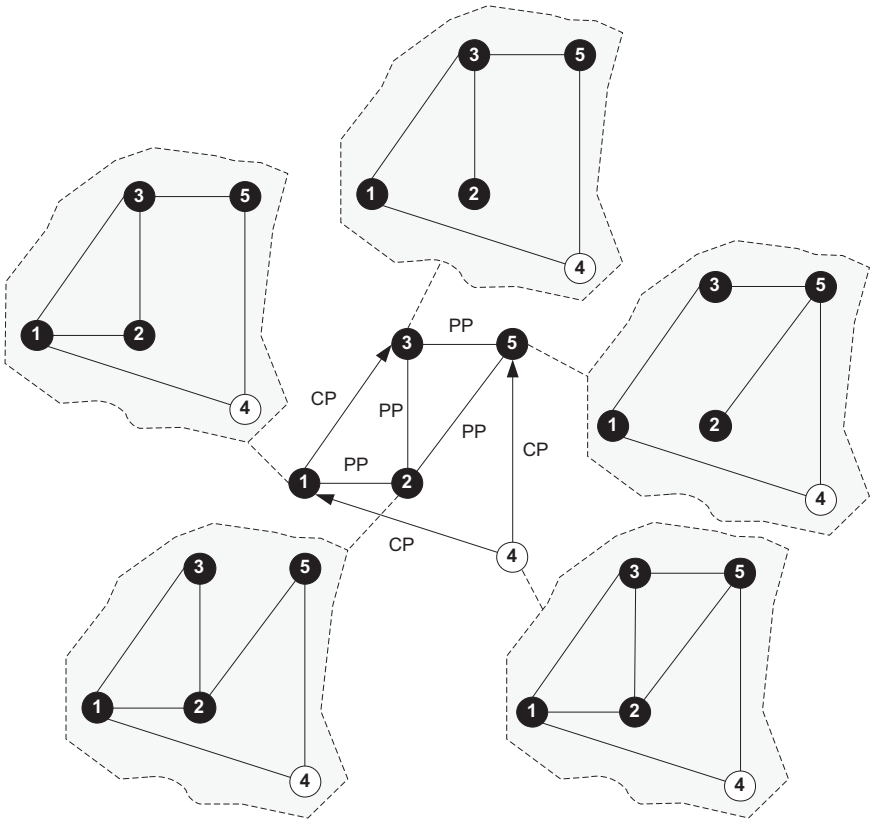

Fig. 5. Inconsistencies in the topological views of domains. These are the topologies that could be potentially inferred by each of domain, according to the valley-free paths available for them.

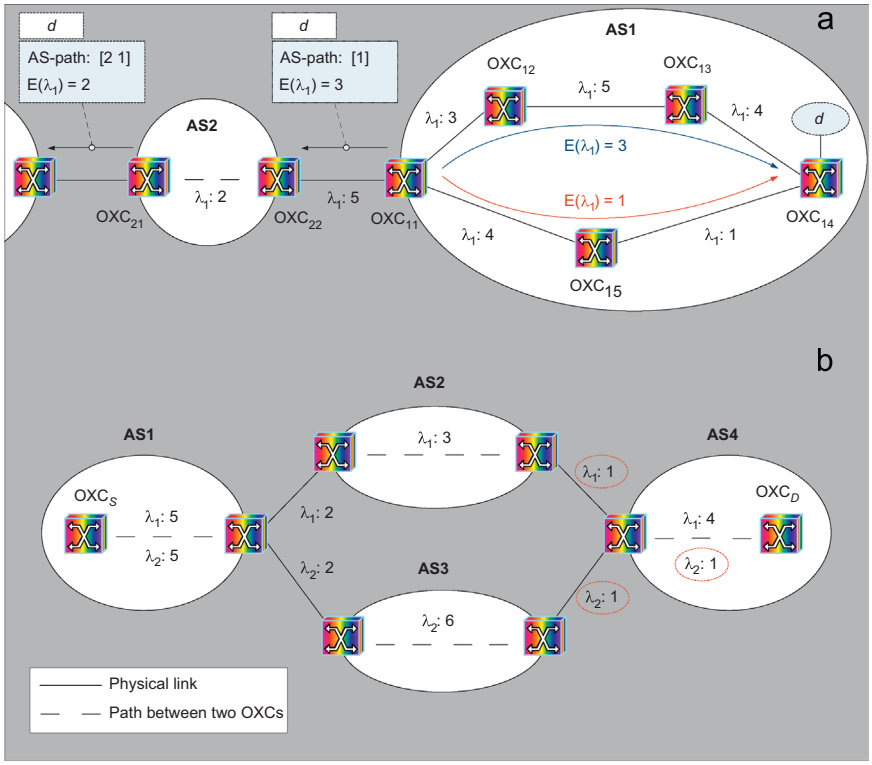

Fig. 6. (a) PVP-1: computation and advertisement of the Effective Number of Available Wavelengths (ENAW) to reach *d*. (b) PVP-2: advantage of computing the load when different paths have the same ENAW.

prefixes and routes can be hijacked. To deepen into the security problems and the threats faced by the routing system the reader is referred to [9].

The effects of routing policies — Routing policies are managed independently and without global coordination among domains. Several studies have shown that, without coordination, the interactions between independent policies may lead to routing anomalies. In fact, there are configurations of routing policies that do not violate any rule in BGP, and despite this, the routing is guaranteed to diverge.

Fig. 5 illustrates an example of the topologies that could be potentially inferred by domains according to the valley-free paths available at each of them. More precisely, due to the valley-free routing policies, the paths available in one domain might not necessarily be known by other domains. For example, even though in the network both AS1 and AS2 are adjacent to AS3 (see the AS graph at the center of Fig. 5), the path [3 5 4] will be

- available for AS1, since this path is valley-free for AS1 (see Table 1);
- unavailable for AS2, since the path is non-valley-free for AS2, and thus it will be unknown for the latter. This is due to the filters applied by AS3 for the transit AS2 $\overset{PP}{\rightarrow}$ AS3 $\overset{PP}{\rightarrow}$ AS5 (see Table 1), as AS3 will not advertise routes learned from AS5 to AS2.

Then, the topologies perceived by the domains in the network are those depicted in Fig. 5. The main conclusion is that the design of routing strategies involving the topology of the network—or abstractions of it—is a challenging task with policy-based path vectors, since the topological views of domains are inconsistent.

Limited Traffic Engineering (TE) control — BGP only offers a limited set of TE functionalities, whose effects are rarely predictable beyond the local domain. Basic TE requirements, such as route control, remain unsolved in practice [11]. Moreover, each BGP router only advertises its best path toward a destination, i.e., the path contained in its FIB, which the one used by the router to forward traffic to the destination. Clearly, this approach improves the overall scalability of the routing system, but adversely reduces the number of paths that can be used for improving the performance and reliability of inter-domain traffic. The lack of effective inter-domain TE tools to control the routing is one of the key missing pieces in the path vector routing model. A particular case of this lack of TE in BGP is the absence of interchange of physical information across domains. BGP can select a inter-domain path which could suffer such physical degradation due to the physical impairments that makes it illegible at destination. To avoid these problems, a inter-domain routing protocol should take into account the quality of the lightpath when selecting the best path.

3. Inter-domain routing in optical networks

Despite the efforts undertaken by the scientific and industrial communities, the issues described in the previous section remain largely unsolved. Indeed, most of the initiatives coming from the research branch underline the

need to find a replacement to the BGP protocol, while providers remain cautious about the issue. In this scenario, the advances made in new switching technologies have opened new frontiers, where it is possible to envision the peering, routing, and switching between domains at different layers, and even in a cross-layer fashion [3]. The cautiousness of providers along with the absence of an alternative routing protocol has led to the proposal of extensions to the BGP protocol specifically developed for certain switching technologies. This section briefly describes a particular BGP extension, the so-called OBGP, and also introduces an alternative approach based on extending a path vector approach with optical extensions.

3.1. The optical border gateway protocol (OBGP)

OBGP is the optical extension of the BGP protocol, and was devised for supporting both the routing and provisioning of lightpaths across domains [2]. In OBGP, the "optical" Network Reachability Information (NRI) is encoded using Multiprotocol BGP (MPBGP) extensions and extended communities [2], and basically consists in the availability of lightpaths to the destination. This NRI allows an OBGP speaker to build up a "lightpath RIB" that can be used for provisioning of optical circuits through different domains. Then, basically it works as BGP but exchanging NRI not about the inter-domain routes but about the inter-domain lightpaths (combination of route and wavelength). If OBGP were to be adopted, multi-domain optical networks would benefit from the advantages of BGP, such as the scalability of the routing system, the confidentiality features, and the distributed management of the routing based on routing policies. Unfortunately, its adoption would also move all the well-known problems in BGP into the routing system of future optical networks [8]. This is one of the main reasons why OBGP has not made progress within standardization bodies.

3.2. Improving the OBGP performance

As mentioned earlier in this paper, the current inter-domain routing system is supported by a path vector routing model that was devised to cover two basic needs: (i) the exchange and distribution of network reachability information and (ii) the distributed management and selection of loop-free routes on a very large scale. However, these objectives are far from current needs and requirements in multi-domain routing. Unfortunately, the tentative approach of designing a new routing protocol for multi-domain networks poses complex challenges, as there are no guarantees that a new protocol can outperform BGP in all the aspects raised in Section 2, and at the same time scale as BGP does. In this section, we shall show that minor modifications to a path vector routing protocol for optical networks can produce significant improvements in terms of performance, and that these modifications can be introduced without impacting on the scalability of the protocol. It has been shown that path vector routing can be considerably improved when path vectors are tagged with *Path-State Information* (PSI) [16,17]. Yet, the integration of PSI without impacting on key aspects of a routing protocol, such as the scalability, the convergence properties, and the number of routing updates, is a challenging task. In this section, we describe a set of very simple modifications to path vector routing in the context of optical networks, mainly focused on the computation and tagging of highly aggregated PSI in the path vectors advertised by optical domains, including (i) intra-domain PSI; (ii) the state of inter-domain links toward downstream domains and; (iii) the already aggregated PSI contained in the inter-domain advertisements received from downstream domains. More specifically, we will briefly describe the operation of three different path vector protocols, namely, PVP-1, PVP-2, and PVP-2-WC, each of which builds upon the previous one. In PVP-1 and PVP-2, the PSIs associated with a lightpath[3] are abstracted as one or two respectively integer values, which are tagged to the lightpath and distributed within the Network Reachability Information (NRI). In [16,17], the authors have shown that the update of PSI can be made without re-advertising NRI. This can be achieved by taking advantage of the *Keepalive* messages exchanged between neighbors. In BGP, Keepalive messages are of fixed length, and consist of only the 19-byte BGP header the PSI. The main advantage of this strategy is that it does not increase the number of routing messages exchanged between optical domains. PVP-2-WC, on the other hand, extends PVP-2 by adding wavelength conversion capabilities at the boundary nodes of optical domains.

3.2.1. PVP-1: aggregated wavelength availability info

The PSI in PVP-1 is composed of aggregated wavelength availability information. To illustrate how this information can be computed and disseminated consider the example in Fig. 6a. In AS1 there are two candidate paths between the border node OXC_{11} and the internal node OXC_{14}. From the figure, notice that for the path through OXC_{12}, the number of wavelengths λ_1 that can be effectively used between OXC_{11} and OXC_{14} is 3, since at most 3 lightpaths can be established without experiencing blocking when λ_1 is assigned. Similarly, for the path through OXC_{15}, the number of wavelengths λ_1 that can be effectively used between OXC_{11} and OXC_{14} is 1. As shown in [17], a simple approach to compute the *Effective Number of Available Wavelengths* (ENAW) of the type λ_1 between OXC_{11} and OXC_{14} is to take the maximum between the two, i.e., $E_{11,14}(\lambda_1)=3$. Fig. 6a shows that with this scheme, the routing protocol PVP-1 in AS1 can advertise upstream (to AS2) a lightpath toward d of the form ([1], λ_1), i.e., with path vector $\overrightarrow{p_1}=[1]$ and wavelength λ_1, and tagged with an ENAW for wavelength λ_1, $E_{11,d}(\lambda_1)=3$.

Algorithm 1. PVP-1 decision process.

Require: $\{P(s,d)\}$ – set of candidate paths between nodes s and d
λ_i – a particular wavelength on the path $P(s,d)$
$E(\lambda_i)$ – the ENAW of wavelength λ_i along the path $P(s,d)$
Ensure: $(P^{best},\lambda^{best})$ – The best lightpath between s and d
1: Choose the (path, wavelength) pair with the highest local preference

[3] A lightpath is represented as a path vector/wavelength pair $(\vec{p},\lambda)$.

2: If the local preferences are equal, choose the shortest AS-path and assign the wavelength with the highest ENAW among the ones available on that path. If more than one wavelength has the same (highest) ENAW along the shortest AS-path, choose the wavelength with the lowest identifier i
3: If the AS-path lengths are equal choose the (path, wavelength) pair associated with the highest ENAW
4: If the ENAWs are equal prefer external paths over internal paths
5: If the paths are still equal prefer the one with the highest ENAW to the next-hop OXC (i.e., to OXC_{n_b} in the neighbor domain)
6: If more than one path is still available run tie-breaking rules

More formally, let u and v be a pair of OXCs inside a domain, where $P(u,v)$ represents a candidate path between u and v, and l, a link within path $P(u,v)$. The routing protocol computes the ENAW of wavelength λ_i between u and v as follows:

$$E_{u,v}(\lambda_i) = \max_{P(u,v)}\{\min_{l\in P(u,v)} E_l(\lambda_i)\} \quad (1)$$

Once the PSI flows outside the domain, the neighbors need to process and possibly update the ENAW aggregated upstream. To compute the inter-domain part of the ENAW, PVP-1 considers the unused wavelengths on the directly connected inter-domain links, and the wavelengths that are available downstream. In concrete, PVP-1 advertises upstream that the ENAW between the local border node OXC_{l_b} and the destination d is

$$E^{adv}_{l_b,d}(\lambda_i) = \min\{E_{l_b,l'_b}(\lambda_i), E'_{l_b,n_b}(\lambda_i), E^{adv}_{n_b,d}(\lambda_i)\} \quad (2)$$

with $E_{l_b,l'_b}(\lambda_i)$ being the ENAW of wavelength type λ_i between two local border nodes, OXC_{l_b} and $OXC_{l'_b}$,; $E_{l'_b,n_b}(\lambda_i)$ the number of free wavelengths of type λ_i in the inter-domain link between $OXC_{l'_b}$ and OXC_{n_b}; and $E^{adv}_{n_b,d}(\lambda_i)$ the ENAW of type λ_i between OXC_{n_b} and the destination d. In the case of Fig. 6a, PVP-1 will advertise upstream $E^{adv}_{21,d}(\lambda_1) = \min\{E_{21,22}(\lambda_1), E_{22,11}(\lambda_1), E^{adv}_{11,d}(\lambda_1)\} = \min\{2,5,3\} = 2$. Algorithm 1 shows a simplified version of the PVP-1 decision process, and how PVP-1 exploits the ENAWs tagged to the candidate lightpaths to make its selection.

3.2.2. PVP-2: aggregated load information

PVP-2 extends PVP-1, where in addition to the ENAW, the routing protocol integrates aggregated load information. This load is captured in the form of a cost function that is associated with each candidate lightpath. The motivation behind the introduction of a load component can be explained through the example shown in Fig. 6(b). In the example, the node OXC_S in AS1 can reach OXC_D in AS4 both through AS2 and AS3, both paths with ENAW 1. This means that the lightpath selection will be almost random in practice, where, in fact, OXC_S should clearly choose the lightpath ([2 4], λ_1), because λ_1 is less "loaded" than λ_2. To address this issue, a cost (an integer number) is tagged to the lightpaths in addition to the ENAW. The cost associated with a candidate path $P(s,d)$ between a local OXC s and a distant OXC d for wavelength type λ_i is computed in PVP-2 as

$$C^{(\lambda_i)}_{P(s,d)} = \begin{cases} \left[\dfrac{1}{E_{s,l'_b}(\lambda_i)} + \dfrac{1}{E_{l'_b,n_b}(\lambda_i)} + \dfrac{C^{(\lambda_i)}_{P(n_b,d)}}{H^{adv}}\right] H \\ \infty \quad \text{if } E_{s,l'_b}(\lambda_i) = 0 \text{ or } E_{l'_b,n_b}(\lambda_i) = 0 \end{cases} \quad (3)$$

In Eq. (3), H represents the number of hops from s to d considering each AS as just one hop. The terms $C^{(\lambda_i)}_{P(n_b,d)}$ and H^{adv} denote the cost and the number of hops, respectively, between OXC_{n_b} and the destination d, advertised by the downstream PVP-2 speaker. Notice that when a wavelength λ_i is unavailable along the path $P(s,d)$, the cost is set to ∞. On this basis, PVP-2 will choose the lightpath ($P(s,d)$, λ_i) with the minimum cost which offers a reasonable trade-off between the length and the wavelength load on the lightpaths chosen. It is worth highlighting that different paths offering the same ENAW will frequently have different costs (loads). A simplified version of the lightpath selection process in PVP-2 would be very similar to that shown in Algorithm 1 but PVP-2 would select first the lightpath with minimum cost, and then would follow the steps in Algorithm 1. In summary, PVP-1 or even PVP-2 introduces three minor changes to legacy path vector routing protocols:

1. The computation and tagging of the ENAWs (and also the costs in PVP-2) to the lightpaths advertised among domains
2. A RWA algorithm exploiting the ENAWs (and the costs in PVP-2).
3. An extended Keepalive message in order to piggyback the updates of both the ENAWs and costs (two integers).

3.2.3. PVP-2-WC: wavelength aggregation with wavelength converters

In order to increase the amount of traffic carried in the network, optical networks can make use of wavelength converters, and thereby increase the availability of wavelengths since the wavelength continuity constraint can be relaxed. A reasonable strategy in multi-domain optical networks is to place wavelength converters at the boundary nodes of optical domains. This is because boundary nodes typically carry large amounts of traffic, hence putting wavelength converters at these nodes may produce significant performance improvements [1]. In concrete, the only difference between PVP-2 and PVP-2-WC is the way in which the ENAW is computed and updated among domains.

4. Modeling physical impairments in a multi-domain optical network scenario

In the previous section we have detailed three RWA protocols using different PSI (*Path State Information*). In these algorithms, only PSI information related to the network state is disseminated and then utilized. However in long optical connections the quality of the optical signal can fall under undesirable values due to the physical impairments, making the signal unreadable at destination.

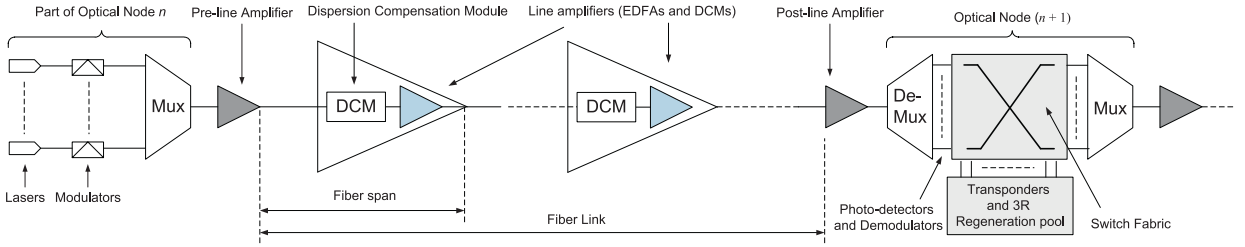

Fig. 7. Main optical elements in an optical transmission system between two adjacent nodes; direction: from node n to node $(n+1)$. Pool of regenerators is optional and their use would break the optical transparency.

This is especially true in inter-domain routes because of their length. In this scenario, the success of a inter-domain connection does not only depend on the PSI utilized in the inter-domain process but also it will depend on the physical information disseminated and used by the routing process.

There exist multiple recent works in the literature addressing the inclusion of physical impairments in the RWA process inside a domain. Two surveys about the main contributions in routing taking into account physical impairments can be found in [18,27]. Nevertheless, very few proposals consider the establishment of a completely transparent lightpath across multiple domains, see [28,29], but none of them succeed on drawing a complete scenario including the most significant physical impairments. Instead, some of the proposals consider that in the border nodes between two domains managed by two distinct operators the optical signal is completely regenerated. For instance, in [19], authors propose a translucent model for impairment aware routing between multiple domains. In this case, lightpaths inside a domain are completely transparent, but the optical signal is regenerated at border nodes, what unfortunately might not be desirable. In fact, the possibility of deploying transparent lightpaths beyond the frontier between domains is gradually being supported by carriers, and thus is attracting increasing attention from the scientific community. Aligned to this, authors in [30] propose to extend the model presented in [19] by aggregating the physical information related to a domain to be interchanged between domains. The physical model in [30], also shown in Fig. 7,[4] is based on transparent optical networks where traffic is transmitted entirely in the optical plane, i.e., without undergoing any optical-electronic-optical (O/E/O) conversion at transit nodes, what extremely reduces the OPEX and CAPEX as well as the energy costs [20].

Next subsections will first describe the physical impairments model within a domain, and afterwards describe how the aggregated information is handled between domains.

4.1. Physical-layer impairment model inside an optical domain

Authors in [30] presented a physical model taking into account the linear and non-linear impairments inside a domain, clearly showing that the physical impairments introduced by the different elements presented in a optical transmission system can produce the following two effects: the degradation of the Optical Signal-to-Noise Ratio (OSNR) and a temporal dispersion of the optical signal. Furthermore, these effects can be linear or non-linear, what substantially affects the way physical impairments are modeled, mainly motivated by the fact that in the case of modeling non-linear impairments, the noise power/dispersion produced by the impairment on a fixed channel (wavelength) depends on the utilization and power levels of the other channels (wavelengths). The model in [30] considers Attenuation, Amplified Spontaneous Emission (ASE), Chromatic Dispersion (CD), and Polarization Mode Dispersion (PMD) as the linear impairments, while node crosstalk and fiber crosstalk (Four Wave Mixing, FWM) are those considered as non-linear impairments all to produce the physical model. The goal in [30] is the aggregation of the effects of the different physical impairments affecting a lightpath turning into two values, noise and delay. Taking into account the linear and non-linear impairments considered above, the Attenuation, the ASE noise, the node crosstalk and the FWM all produce a OSNR degradation and they were modeled as a noise, whereas CD and PMD produce different group velocities of the spectral components and thus producing a delay.

In summary, the whole effects of the different elements in an optical transmission system are an added noise, that is OSNR, and a dispersion, both affecting the optical signal. Indeed, from now on, we consider $o_{x'x}^{(\lambda_l)}$ and $\Delta t_{x'x}^{(\lambda_l)}$ as the OSNR and the dispersion respectively introduced by the physical impairments between nodes x' and x $(x' \rightarrow x)$ for the wavelength (λ_l). We assume that nodes x' and x are inside the same domain and $o_{x'x}^{(\lambda_l)}$ and $\Delta t_{x'x}^{(\lambda_l)}$ can be computed by the source node or a centralized node inside the domain.

We now export these impairments aggregation model to a multi-domain scenario, clearly describing its use in the frontier between two neighbor domains.

4.2. Information exchange model between adjacent optical domains

Consider the scenario shown in Fig. 8, wherein two adjacent domains, AS_{i-1} and AS_i, have an agreement to foster the set up of transparent optical circuits between them. The goal is to provide an information exchange model enabling optical bypass at the frontier between AS_{i-1} and AS_i, while preserving the confidentiality of the physical-layer information managed by each domain.

[4] As shown on the right-hand side of Fig. 7, a pool of regenerators can be optionally connected at dedicated ports of the switching fabric, but its use would clearly break the optical transparency.

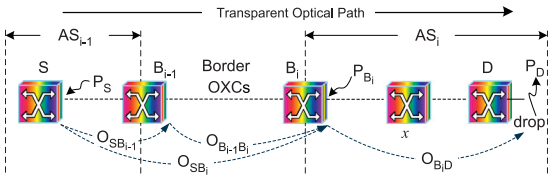

Fig. 8. Optical bypass between two neighbor domains.

To this end, let $S \in AS_{i-1}$ be the source node, and B_i and D be the ingress (border) node and the destination node in AS_i, respectively (see Fig. 8). Let $O_{x'x}$ and $\Delta T_{x'x}$ be two vectors of size W (number of wavelengths) that we call the OSNR and dispersion vectors for the segment of a path between the nodes x' and x ($x' \rightarrow x$), which we define as follows:

$$O_{x'x} = \begin{bmatrix} o_{x'x}^{(\lambda_1)} \\ \vdots \\ o_{x'x}^{(\lambda_l)} \\ \vdots \\ o_{x'x}^{(\lambda_W)} \end{bmatrix}, \quad \Delta T_{x'x} = \begin{bmatrix} \Delta t_{x'x}^{(\lambda_1)} \\ \vdots \\ \Delta t_{x'x}^{(\lambda_l)} \\ \vdots \\ \Delta t_{x'x}^{(\lambda_W)} \end{bmatrix} \tag{4}$$

when x represents a node inside AS_i, and $x' = B_i$, i.e., an ingress node to AS_i. According to the aggregation strategy defined in the past subsection, the term $O_{x'x}$ represents the relation signal to noise due to the attenuation, the ASE noise, the node crosstalk and the FWM noise. The second term, $\Delta T_{x'x}$, aggregates all the accumulated delay due to Chromatic Dispersion (CD) and Polarization Mode Dispersion. Furthermore, let $o_{\min}$ and $\Delta t_{\max}$ be the minimum acceptable OSNR and the maximum admissible dispersion at detection, respectively, so that a lightpath can be established between the source and destination—note that these bounds are technology dependent, and thus will vary depending on the nodes and fibers used within an optical domain. Assuming that the admission control policy in AS_i is satisfied, then, the set up of a new lightpath between S and D through B_i could be provisioned transparently using wavelength λ_l when

$$\begin{cases} o_{SD}^{(\lambda_l)} = \left(\underbrace{\frac{1}{o_{SB_{i-1}}^{(\lambda_l)}} + \frac{1}{o_{B_{i-1}B_i}^{(\lambda_l)}}}_{(o_{SB_i}^{(\lambda_l)})^{-1}} + \frac{1}{o_{B_iD}^{(\lambda_l)}} \right)^{-1} > o_{\min} \\ \Delta t_{SD}^{(\lambda_l)} < \Delta t_{\max} \end{cases} \tag{5}$$

Clearly, both inequalities must be satisfied in order to allow optical bypass at the border node B_i in AS_i. It is also worth noting that in the first inequality in (5):

- $o_{SB_{i-1}}^{(\lambda_l)}$ can be estimated by AS_{i-1}, and thus it can be used by an impairment-aware RWA algorithm prior to the path setup process;
- a nominal value characterizing the OSNR for the inter-domain link ($o_{B_{i-1}B_i}$) can be either advertised between AS_{i-1} and AS_i or estimated prior to the path setup process.

Thus, the terms $o_{SB_i}^{(\lambda_l)}$ and the dispersion in (5) between S and B_i can be reasonably estimated by AS_{i-1}. On this basis, we propose to extend the budget-based approach introduced by Yang et al. [21] as follows. By operating in (5) we obtain the inequalities that AS_{i-1} must satisfy so that AS_i allows optical bypass in B_i

$$\begin{cases} o_{SB_i}^{(\lambda_l)} > \dfrac{o_{\min}}{1 - o_{\min} \cdot (o_{B_iD}^{(\lambda_l)})^{-1}} = o_{\mathrm{Budget}}^{(\lambda_l)} \\ \Delta t_{SB_i}^{(\lambda_l)} < \Delta t_{\max} - \Delta t_{B_iD}^{(\lambda_l)} = \Delta t_{\mathrm{Budget}}^{(\lambda_l)} \end{cases} \tag{6}$$

where $o_{B_iD}^{(\lambda_l)}$ and $\Delta t_{B_iD}^{(\lambda_l)}$ are computed by AS_i. As in (4), we define two budget vectors of size W, namely, O_{Budget} and $\Delta t_{\mathrm{Budget}}$, such that each component of these vectors is given by the right-hand side of the first and second inequalities in (6), respectively. In this framework, the budget vectors O_{Budget} and $\Delta t_{\mathrm{Budget}}$ represent the information that AS_i will associate to destinations D within its own domain, guaranteeing transparent transit. Accordingly, the budget vectors are the information that AS_i will advertise to AS_{i-1}. In case that the local OSNR in AS_i is below the minimum admissible OSNR, $o_{\min}$, or the local dispersion surpasses $\Delta t_{\max}$ for a given (path, wavelength, destination) tuple, the budgets advertised for the tuple are set to infinite, and zero, respectively. It is important to note that, in our information exchange model, detailed physical-layer information is never disclosed between AS_i and AS_{i-1}, since only two scalar values are exchanged per (path, wavelength) pair toward any given destination D.

5. Modifying the control plane to accommodate physical impairments

In this section we will introduce the modifications required in the optical control plane to compute inter-domain routes with TE capabilities, including physical impairments as modeled in Section 4. Before going into the multi-domain scenario we review some of the recent works proposing control plane extensions required to disseminate physical information inside a domain.

5.1. Control plane extensions to consider physical impairments: intra-domain case

We first mention different proposals for extending the control plane to support physical impairments in a intra-domain scenario. Basically we can divide these proposals into three different classes: (i) those extending OSPF-TE to disseminate physical information among the nodes of the domain [24], so-called the routing approach; (ii) those proposing changes in the RSVP-TE protocol, such as [24,25], aiming at considering the physical impairments during the setup process, so-called the signaling approach, and (iii) those so-called the probing approach. While in the first case, the routing approach, physical impairments are considered by the nodes in the domain when computing a lightpath, they are not in the second case, the signaling approach. In the latter, the physical impairments are checked node by node during the lightpath setup process. There exist also the option of combining both, the routing approach and the signaling approach. In this mixed scenario OSPF-TE extensions are used to disseminate linear physical impairments, whereas the RSVP-TE protocol is

extended to check the non-linear physical impairments. Detailed studies of the different cases and their implications in terms of control plane extensions can be found in [27]. Finally in the third case, the probing approach, probe traffic is injected in a lightpath in order to measure its Bit Error Rate (BER) at the destination node to check for the physical availability. An illustrative example of a probe-base solutions can be found in [26].

5.2. Control plane extensions to consider physical impairments: inter-domain case

As it is stated above, most of the proposals addressing multi-domain routing considers optical signal regeneration in the border nodes [19]. In fact, only a few proposals deal with the establishment of a completely transparent inter-domain path. A line of work proposes to compute the lightpath in a backward way starting from the destination node, and then checking the physical impairments being accumulated by the route in each domain. This is the case of [28], where authors propose a multi-layer and multi-domain impairment aware RWA (IA-RWA) routing algorithm based on PCE computation, that also considers OSPF-TE extensions to disseminate physical information within domains. Two main drawbacks constraint a potential deployment of this proposal. On one hand, the long delay in the setup process inherent to the backward strategy where the physical information is computed online for each PCE-node in each domain. On the other hand, this proposal only considers OSNR (attenuation), crosstalk and PMD as physical impairments, while skipping others that are also relevant.

Another contribution comes from [29] where authors address the recent work done in standardization bodies to endow PCE with capabilities for multi-domain and impairment aware routing. But, the work is not explicitly putting all set of constraints together, that is, extensions to the PCE for multi-domain routing taking into account the physical impairments. In the concrete case of the required PCE extensions to deal with physical impairments, authors support their proposal through a previous work [23], where they propose a PCE-based solution that takes into account physical parameters during the path computation. To this end, a physical parameter database (PPD) collecting physical information is included in the PCE architecture and is updated by means of OSPF-TE extensions.

So far, we have analyzed two different lines of work, one based on a backward strategy and another one based on decoupling functionalities through PCE. A different option focuses on source routing. In this case, the source domain is responsible for computing a completely transparent lightpath across domains based on the aggregated physical information being exchanged between domains.

5.3. Control plane extensions based on IDRAS

As we have shown in the previous sections, a multi-domain routing model mostly centered on the exchange of network reachability information (NRI), like the one we currently have with BGP or the one offered by OBGP, will not be sufficient to meet the specific requirements of multi-domain optical routing. We have also shown that, in addition to NRI, other information must be interchanged between neighboring domains, such as aggregated PSI enriched with physical impairments advertised through for example the budget vectors defined in Section 4.2. In this section we review a distributed control plane strategy proposed in [16] which efficiently exploits the aggregated PSI in a multi-domain setting. The key point in this control plane strategy is that it may be extended to offer a simple way to accommodate physical information advertisements in the form of budget vectors. This extension is illustrated in this section.

In the control layer each AS will be a routing control domain (RCD) that may allocate one or more inter-domain routing agents (IDRAs), depending on its scale (see Fig. 9). The role of the IDRAs is twofold. On one hand, they are the ones that distribute the routing and signaling information between RCDs. On the other hand, they are in charge of the computation and establishment of inter-domain lightpaths in a distributed way similar to the path computation element (PCE) model [22].

The establishment of a inter-domain lightpath is performed by the IDRAs in three phases: routing, signaling, and setup. During the routing phase, the IDRA in the source domain uses the information advertised by neighboring IDRAs to find a loose end to-end lightpath between the source and the destination node. The advertisements distributed by the IDRAs contain the usual NRI, in addition to TE information consisting of PSI and the set of offered services by the RCDs along a path (see Fig. 9). During the composition of the advertisements, the IDRAs aggregate the PSI along a path, taking into account the state of both the intra-domain and inter-domain segments of the path. Then the information distributed by the IDRAs will be

- Network Reachability Information (NRI). Conversely to BGP, the NRI exchanged among the IDRAs does not include the AS-path to reach a destination. The IDRAs use the TE information contained in the routing

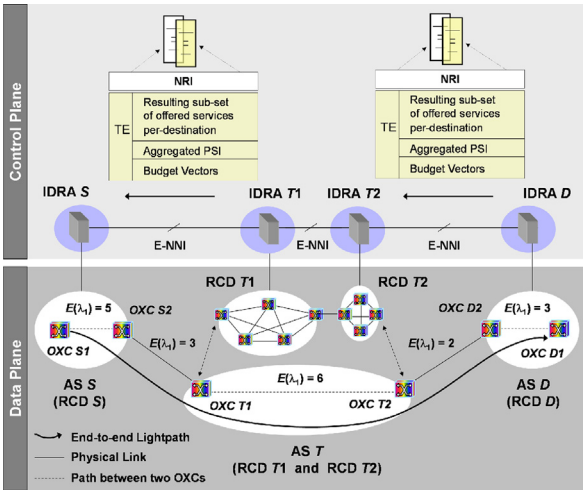

Fig. 9. Architecture of the IDRA-based routing and TE control model.

advertisements to compare the routes. Another important difference is that the IDRAs can advertise multiple routes per destination, even with the same NH address. The NRI information is composed by:
  - Set of destinations.
  - Next Hop of these destinations (NH).
  - A set of pairs (λ_i, M λ_i) for each destination, where λ_i denotes a particular wavelength i and M λ_i denotes the maximum multiplicity advertised for λ_i.
- Path State Information (PSI). The IDRAs advertise PSI messages by aggregating and assembling the following three pieces of information: (i) intra-domain PSI; (ii) PSI related to the inter-domain links towards its downstream domains; and (iii) the already aggregated PSI contained in the inter-domain advertisements received from downstream domains. The PSI information that the IDRA of domain AS_i advertises to IDRA of domain AS_{i-1} is
  - Aggregated Wavelength Availability Information (ENAW): The computation of the ENAW is a simple process, where an IDRA first keeps the minimum number of available wavelengths on the links of a candidate path, and then computes the maximum among all candidates.
  - Aggregated Load Information (Cost): The goal is that the cost reflects the current load in the availability of wavelengths in a inter-domain path, allowing an IDRA to compare routes more accurately than directly using the ENAWs between the source and destination.
    So far, PSI messages do not embed physical impairments information. But, due to its open nature including this information can be easily done by simply adding the budget vectors in the PSI, as follows:
  - Aggregated Physical Information (Budget Vectors): According to Eq. (6) the budget vectors O_{Budget} and $\Delta t_{\mathrm{Budget}}$ are computed by the IDRAs for every destination and they represent the physical information that the IDRA associates to every possible destination, inside or outside the domain.

In summary, the PSI advertised by the IDRAs consists of a set of candidate lightpaths, together with their ENAWs, costs and budget vectors. In [16] we proposed to extend the Keepalive messages of BGP with the purpose of piggybacking ENAWs and costs, only when relevant information needs to be updated. Now, we also propose to include the budget vectors in these Keepalive messages. The algorithms PVP-1 and PVP-2 previously reviewed utilize the PSI information (ENAW and Cost) to select the inter-domain lightpaths. These algorithms can be improved to also consider the budget vectors in their decision process. For example with a lightly variation, the PVP-2 routing protocol can consider this physical information; the source IDRA should select the lightpath minimizing the cost between the lightpaths with a $OSNR^{s\to d}$ higher than a minimum OSNR threshold, $O_{\min}$, and with $\Delta t^{s\to d}$ lower than the maximum dispersion allowed, $\Delta t_{\max}$.

6. Multi-layer and multi-domain optical routing

In previous sections we started analyzing solutions for inter-domain routing in optical networks. Then, we extended the scope to also include physical impairments in the decision process in the route control architecture. However, proposals analyzed so far strictly focus on optical networks. In this section we incrementally introduce different concepts from multi-domain to multi-layer in the routing problem.

Current research in backbone networks proposes backbone optical networks with a data plane switching at different granularities with a Generalized Multi-Protocol Label Switching (GMPLS) control plane. For example, a packet switching (IP/MPLS) upper layer over an optical WDM layer switching at the wavelength level. In this scenario, the lower layer acts as a server to the client upper layer. When the routing process in the IP layer does not find any available connection between two IP routers, it can request a new connection to the optical layer (lightpath) creating a new link in the logical topology in the IP layer. Different studies in the literature argue the importance of optimizing globally the resources of a multi-layer network, rather than optimizing every layer independently [31] , this is known as multi-layer traffic engineering (TE).

Several works in the literature propose solutions for multi-layer networks [31,33] and only a few extend the work for multi-layer/multi-domain networks [31,32], most of them based on the Path Computation Element (PCE) [22]. In a completely centralized approach where there would only be a unique PCE for all layers and domains in the network, scalability issues will easily come up. Thus, distributed approaches with multiple PCEs distributed among the layers/domains of the network have been already proposed.

As said before, several contributions mostly addressed the multi-layer network scenario (non-multi-domain). In this context, it must be noticed that the cooperation between PCEs in a multi-layer network can follow either the Horizontal Approach (HA-PCE) [33] or the Vertical Approach (VA-PCE) [34]. In HA-PCE different PCEs are assigned to the different layers in the network, whereas in the VA-PCE one unique PCE controls a subset of nodes belonging to different layers. Despite reducing the information distribution scenario for a single PCE, the HA-PCE approach still suffers from scalability issues due to the excessive exchange of information between the nodes of one layer and their corresponding PCE. The VA-PCE was proposed in [34] to solve the scalability issues in the HA-PCE. This VA-PCE approach can be extended to also consider a multi-domain network scenario. Thus, for multi-layer/multi-domain networks, authors in [32] propose a VA with a PCE for every domain controlling the nodes of the different layers belonging to the same domain. In particular, each PCE has the capability to switch at two different layers, the optical, that is wavelength switching nodes (OXCs) and the network, that is packet switching nodes (IP routers). One of the main characteristics of this proposal is that the PCE considers not only the established lightpaths at the optical layer, but also those not yet established (so-called feasible TE links) when computing the multi-layer path.

So far, we have analyzed main research trends addressing the routing problem in multi-layer/multi-domain optical networks. Following the incremental approach, we now introduce the physical impairments information in the routing decision process. Unlike those proposals described in Section 5, this section deals with a multi-layer network scenario. Few works in the literature address the routing in multi-layer/multi-domain networks considering physical impairments. In [35] authors propose the use of a centralized optical PCE in the optical layer interworking with a Network Management System (NMS) in the upper layer, that takes into account the physical impairments in the path computation. In this paper, OSNR, PMD and aggregated non-linear effects are advertised between the PCE and the NMS. However, this proposal does not consider the computation of multi-domain paths.

We now introduce a novel and simple approach to the routing problem in a multi-domain/multi-layer network scenario with physical impairments, this approach is based on extending the control strategy proposed in Section 5.3 for a multi-domain scenario, to be applied also to a multi-layer scenario. To this end, we propose a novel HA strategy based on integrating the IDRA's functionalities in the PCE at the optical layer, hence generating a multi-layer infrastructure with a PCE (with IDRA functionalities) on each domain in the optical layer and a PCE for each domain in the IP layer, all communicated through the PCE communication protocol (PCEP) [36]. On the other hand, the IDRA's functionalities could be also integrated in a PCE controlling the nodes of both layers of a domain so matching the definition of a Vertical Approach (VA). In Fig. 10, we show an example of VA approach for a multi-layer/multi-domain network. Observe that the physical topology, OXCs and fiber links do not match exactly with the IP topology, IP routers and links. In this example, only the topological information of the optical layer is aggregated by the IDRAs. Fig. 10 shows both the optical network without aggregation and the abstraction layer obtained after topology aggregation. Indeed, Fig. 10 illustrates how AS T1 and AS T2 are mapped into two nodes, OXC T1 and OXC T2. However, in the IP layer only the complete (no aggregation) topology is presented. In this VA, the PCE managing both layers should be able of aggregating also the IP layer information as IDRAs do at the optical layer [37]. This aggregated IP information would be also advertised and utilized by the PCEs to computed inter-domain IP paths.

In Fig. 11 we illustrate the proposed architecture through an example of routing in a multi-layer/multi-domain network. Let us assume that a connection must be established between nodes S, in AS S and node D in AS D, at the IP layer. To this end, the node S will request PCE-1 in AS S for a connection to D. Then, PCE-1 will compute the connection according to the aggregated information it receives from other PCEs in the route to D. This information stands for the aggregated available bandwidth [37], in a similar way the optical availability is aggregated in the form of ENAWs. With this information, PCE-1 will compute the path in the IP layer across the different domains to D. Let us assume that there is real connectivity between node S in AS S and node A in AS 1, but due to some reason (congestion, failure, etc.) there is no IP connectivity from Node A on. In this case it is necessary to request a new lightpath to the optical layer. The IDRA-T1 integrated in PCE-2 will compute and set up a lightpath in the optical layer, between OXC T1-A in AS T1 and OXC D-B in AS D, the dark blue line in the optical layer. This new connection will be updated in the IP topology and it will be available for

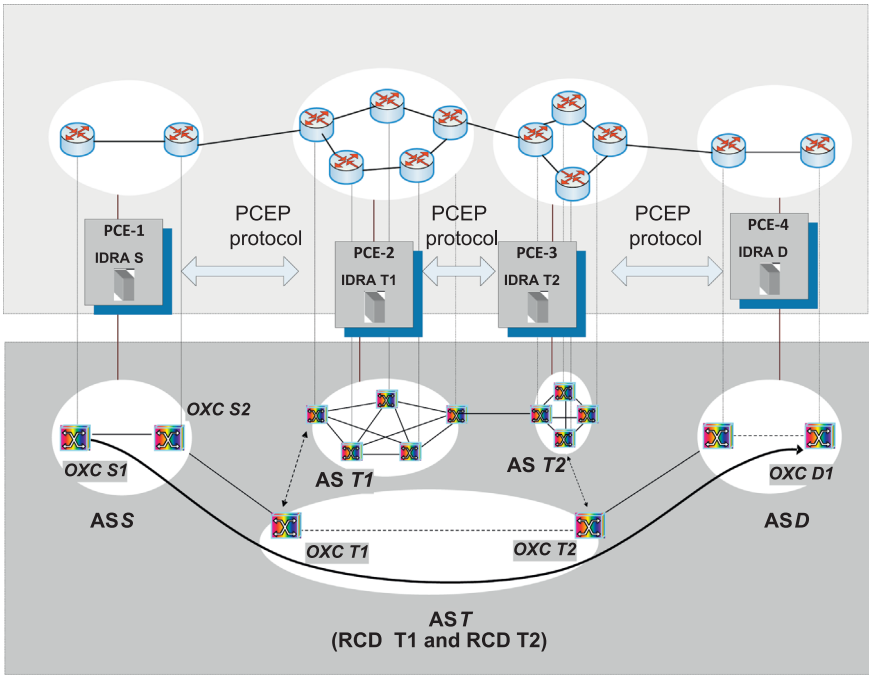

Fig. 10. VA PCE approach for a multi-layer/multi-domain network.

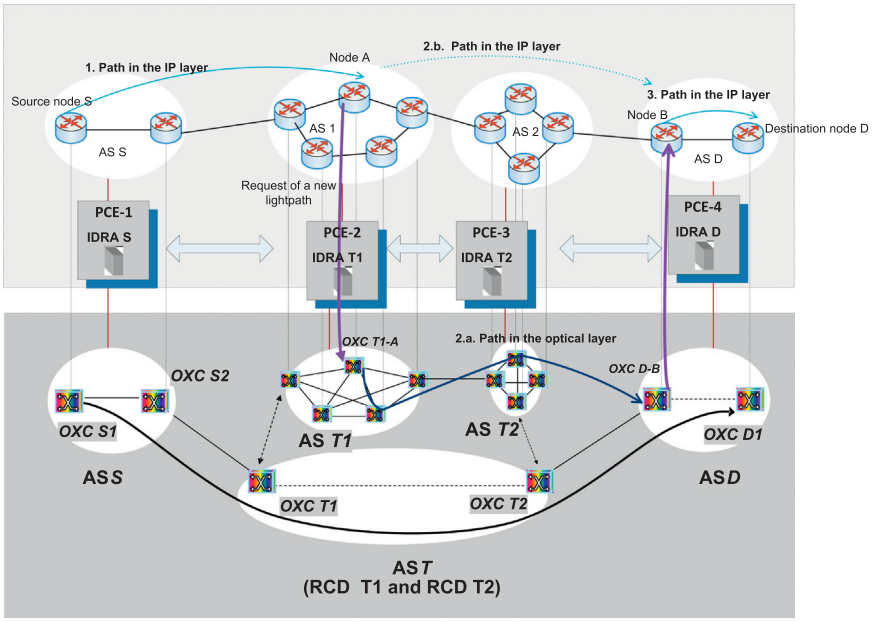

Fig. 11. Example of path computation in a multi-layer/multi-domain network. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

the establishment of an IP connection in the upper layer, the dashed clear blue line in the IP layer. This is a clear example of PCE and IDRA integration on a multi-domain/multi-layer network scenario.

7. Validating the concepts

In this section we compare the performance of the path vector protocols proposed in Section 3, PVP-1, PVP-2 and PVP-2-WC, versus OBGP. To this end, we analyze four different performance metrics: (a) the blocking ratio (cf. Figs. 13 and 15); (b) the overall number of routing messages (cf. Fig. 14); (c); and the convergence time after a node failure (cf. Fig. 16). In the next subsection, we first describe the simulation methodology and then we present the obtained results.

7.1. Evaluation methodology

We have carried out extensive simulations using OPNET Modeler to evaluate the above listed metrics. In these simulations, we use the PAN-European network topology shown in Fig. 12, which has been widely used as a reference optical network topology in several research contributions. The network consists of 28 domains and 41 inter-domain links. Inside each domain, we placed a random number of OXCs equal to or higher than the number of inter-domain links on that domain. For example, the number of OXCs inside Munich is at least 4 for all the scenarios considered in our tests.

The traffic was simulated between different domains, considering 18 sources and 10 destination OXCs randomly located inside the domains covering the entire PAN-European network. Each link in the network consists of 5 fibers and each fiber has 14 wavelengths. After thousands

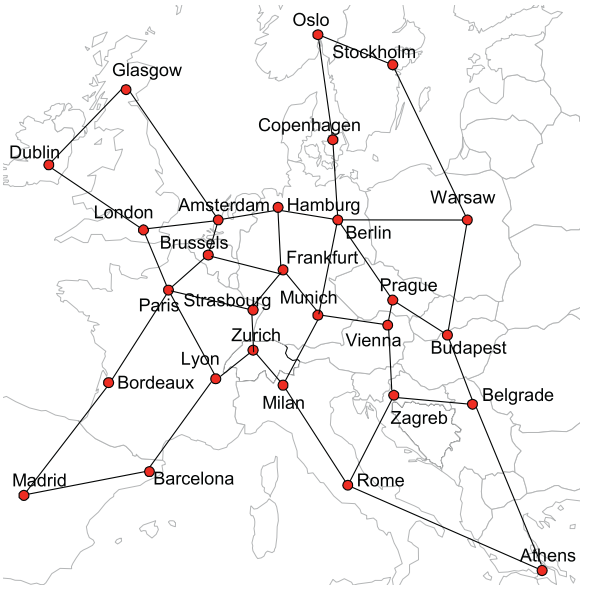

Fig. 12. PAN-European reference network.

of hours of event-driven simulations we have chosen both the number of OXCs and the number of fibers per link, in order to provide a good trade-off between the size of the network and the time needed to run the tests. Traffic was modeled according to a Poisson distribution with exponentially distributed arrival and departure rate, ranging from 100 up to 300 Erlangs. Furthermore, in the obtained results, the frequency of updates within the PSI messages has been normalized to the default *Keepalive Update Interval* currently used in most BGP implementations, which is 60 s.

The presented results are the averages of over 100 randomly generated PAN-European network configurations. To this end, different network configurations are generated by changing both the topology and the source and destination OXCs within each domain.

7.2. Introducing obtained results

Fig. 13 shows the blocking ratio obtained for different traffic loads by the 2 different proposed algorithms versus OBGP. From the results, we observe that both PVP-1 and PVP-2 substantially improve the results obtained by OBGP. While OBGP experiences blocking for all the traffic loads tested, PVP-1 and PVP-2 only show some negligible blocking after reaching 200 Erlangs. Depending on the traffic load, OBGP yields an overall blocking that is approximately between 7 and 350 times larger than the one obtained with PVP-2.

In Fig. 14 the number of routing messages versus the traffic load is plotted. The results confirm that the improvements in the blocking ratio shown in Fig. 13 are obtained without adversely affecting the churn rate of updates. In fact, PVP-1 and PVP-2 always need a smaller overall number of routing messages than OBGP. This is mainly due to two reasons. First, PSI updates are never sent directly between neighbors but rather they are piggybacked in the Keepalive messages exchanged between them. Second, OBGP tends to exhaust the available wavelengths along the shortest AS-path before switching to an alternative path, what unfortunately increases the number of network reachability messages when paths become blocked (producing the path exploration problem as well). Instead, PVP-1 and PVP-2 explicitly consider the ENAW, and PVP-2 also considers the cost, hence they are both able to provide much better traffic distribution than OBGP. This characteristic produces a significant reduction in the blocking ratio, and therefore, less network reachability messages need to be exchanged.

In Fig. 15 we draw the improvements obtained when using wavelength converters in PVP-2. From the figure we observe that PVP-2-WC outperforms PVP-2, which in turn, outperforms both PVP-1 and OBGP. It is worth observing that, for the case of 10 converters, PVP-2-WC achieves a blocking ratio lesser than 0.1% for all traffic loads simulated. This value is the blocking threshold recommended in order to support real-time and streaming applications in optical networks.

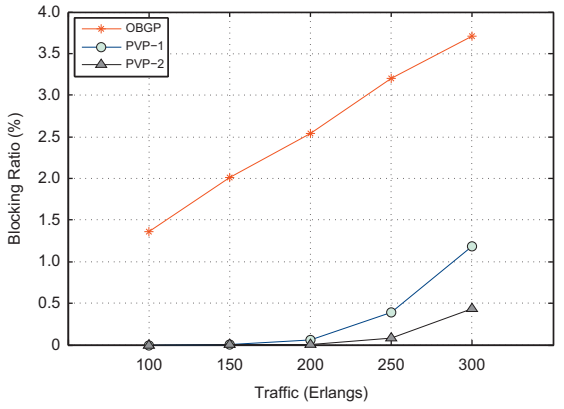

Fig. 13. Average blocking ratio obtained with OBGP, PVP-1, and PVP-2.

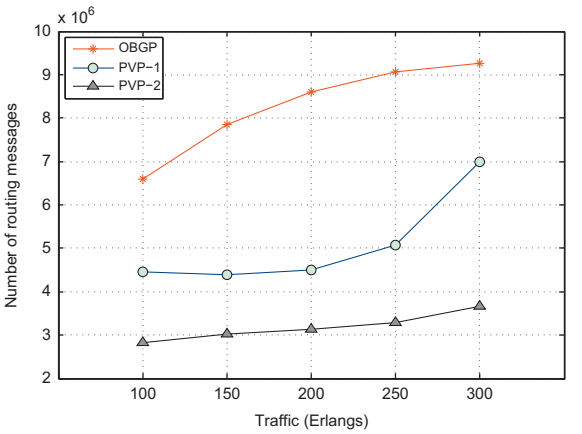

Fig. 14. Number of routing messages obtained with OBGP, PVP-1, and PVP-2.

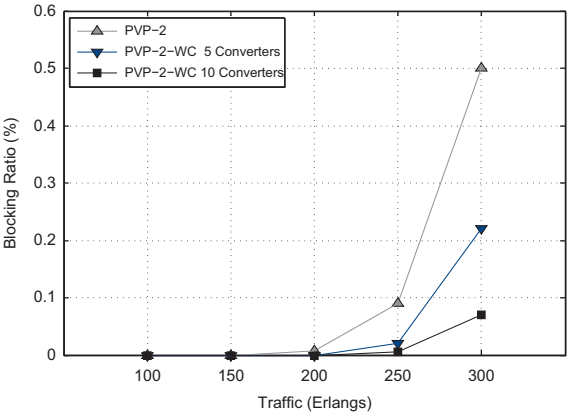

Fig. 15. Average blocking ratio obtained with PVP-2 and PVP-2-WC; comparison using 5 and 10 wavelength converters in PVP-2-WC.

Authors have carried out a large set of experiments and hence obtained a large set of results, those due to the space constraints are not included in this paper. However, from the overall set of obtained results we can conclude that the number of messages generated decreases as more wavelength converters are used in the network. The reason is that in the presence of wavelength converters, more wavelengths are available along the paths, which undoubtedly reduces the blocking, which in turn reduces the exchange of reachability messages and the path exploration significantly.

In order to analyze the convergence properties of the proposed routing algorithms, in the last experiment shown in this paper we compare their performance under stressful conditions. More precisely, we measure the time elapsed between a node failure, and the instant when the last message originated by this event is processed. The assessment of the impact of this kind of event on an inter-domain RWA protocol is particularly important, since

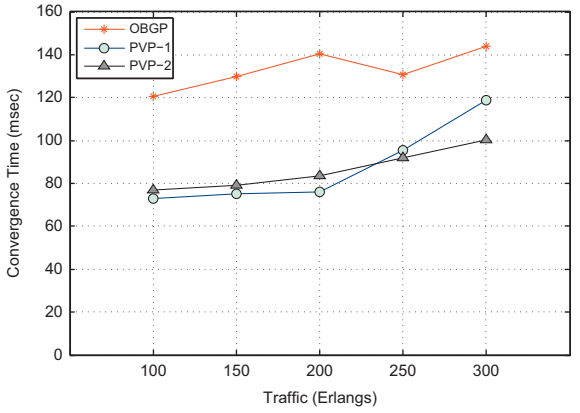

Fig. 16. Convergence time (in milliseconds) during a node failure in Frankfurt.

applying protection and restoration techniques to all the lightpaths across a node might not be feasible in practice.

Fig. 16 shows the convergence time during a node failure. It is important to notice that even for a small multi-domain optical network (see Fig. 12), the time required by an optical path-vector protocol (OBGP) to converge upon a node failure is far large than those demanded by either PVP-1 or PVP-2.

8. Conclusions

This paper presents a detailed analysis of the reasons still pushing for active research in the field of multi-domain routing in optical networks, as well as a review of proposals and research lines through an incremental perspective, including several aspects that strongly impact on the lightpath selection on a multi-domain scenario. We first clearly argue the reasons supporting solutions based on BGP extensions and the most appropriate approaches to address the multi-domain routing problem in optical networks. In this regard, we describe the main BGP limitations and weaknesses and their potential impact on the optical layer. The paper reviews and evaluates different proposals considering the *Path State Information* (PSI) in the routing process on an optical scenario. This network scenario is then extended to a multi-layer/multi-domain scenario where a control architecture based on inter-domain routing agents taking into account the physical impairments and aiming at facilitating the routing process is also proposed.

Acknowledgments

This work was partially funded by the Spanish Ministry of Science and Innovation under contract TEC2009-07041, and the Catalan Research Council under contract 2009SGR1508.

References

[1] A. Beshir, M. Yannuzzi, F. Kuipers, Inter-domain Routing in Optical Networks with Wavelength Converters, in: Proceedings of IEEE ICC, Cape Town, South Africa, 2010.
[2] M. Blanchet, F. Parent, B.St. Arnaud, Optical BGP (OBGP): InterAS Lightpath Provisioning, IETF Draft, 2001.
[3] M. Chamania, A. Jukan, A Survey of inter-domain peering and provisioning solutions for the next-generation optical networks, IEEE Communications Surveys & Tutorials 11 (1) (2009). (First Quarter).
[4] J. Chandrashekar, Z. Duan, Z.L. Zhang, J. Krasky, Limiting path exploration in BGP, in: Proceedings of IEEE INFOCOM, Miami, FL, USA, 2005.
[5] CIDR Report, ⟨http://www.cidr-report.org/as2.0/⟩, February 2013.
[6] T. Griffin, G. Wilfong, An Analysis of BGP Convergence Properties, ACM/SIGCOMM, Cambridge, MA, USA, 1999.
[7] C. Labovitz, A. Ahuja, A. Bose, F. Jahanian, Delayed internet routing convergence, IEEE/ACM Transactions on Networking 9 (3) (2001) 293–306.
[8] X. Masip, M. Yannuzzi, Optical Multi-Domain Routing IEEE/OSA OFC, San Diego, USA, 2009.
[9] M.O. Nicholes, B. Mukherjee, A survey of security techniques for the Border Gateway Protocol (BGP), IEEE Communications Surveys and Tutorials 11 (1) (2009) 52–65.
[10] R. Oliveira, B. Zhang, D. Pei, R. Izhak-Ratzin, L. Zhang, Quantifying path exploration in the internet, IEEE/ACM Transactions on Networking 17 (2) (2009) 445–458.
[11] B. Quoitin, BGP-based Interdomain Traffic Engineering, Doctoral Thesis, Louvain-la-Neuve, Belgium, 2006.
[12] B. Quoitin, S. Uhlig, C. Pelsser, L. Swinner, O. Bonaventure, Inter-domain traffic engineering with BGP, IEEE Communications Magazine 41 (5) (2003) 122–128.
[13] Y. Rekhter, T. Li, S. Hares, A Border Gateway Protocol 4 (BGP-4), IETF RFC 4271, 2006.
[14] S. Secci, J.L. Rougier, A. Pattavina, F. Patrone, G. Maier, PEMP: Peering Equilibrium Multipath Routing, IEEE GLOBECOM, HI, USA, 2009.
[15] A. Yahaya, T. Harks, T. Suda, iREX Efficient automation architecture for the deployment of inter-domain QoS policy, IEEE Transactions on Network and Service Management 5 (1) (2008) 50–64.
[16] M. Yannuzzi, X. Masip-Bruin, G. Fabregó, S. Sánchez-López, A. Sprintson, A. Orda, Toward a new route control model for multidomain optical networks, IEEE Communications Magazine 46 (6) (2008) 104–111.
[17] M. Yannuzzi, X. Masip-Bruin, S. Sánchez-López, E. Marín-Tordera, OBGP+: an improved path-vector protocol for multi-domain optical networks, Optical Switching and Networking 6 (2) (2009) 111–119.
[18] S. Azodolmolky, M. Klinkowski, E. Marin-Tordera, D. Careglio J. Sole-Pareta, I. Tomkos, A survey on physical layer impairments aware routing and wavelength assignment algorithms in optical networks, Computer Networks 53 (7) (2009) 926–944.
[19] M. Gagnaire, S. Al Zahr, Impairment-aware routing and wavelength assignment in translucent networks: state of the art, IEEE Communications Magazine 47 (5) (2009) 55–61.
[20] R.S. Tucker, R. Parthiban, J. Baliga, K. Hinton, R. Ayre, W.V. Sorin, Evolution of WDM optical IP networks: a cost and energy perspective, Journal of Lightwave Technology 27 (3) (2009) 243–252.
[21] X. Yang, B. Ramamurthy, Dynamic routing in translucent WDM optical networks: the intradomain case, IEEE Journal of Lightwave Technology 23 (3) (2005) 955–971.
[22] A. Farrel, J.P. Vasseur, J. Ash, A Path Computation Element (PCE)-Based Architecture IETF RFC 4655, 2006.
[23] F. Cugini, F. Paolucci, L. Valcarenghi, P. Castoldi, Implementing a path computation element (PCE) to encompass physical impairments in transparent networks, in: Proceedings of IEEE OFC/NFOEC, Anaheim, CA, USA, 2007.
[24] Ricardo Martínez, Carolina Pinart, Challenges and requirements for introducing impairment-awareness into the management and control planes of ASON/GMPLS WDM networks, IEEE Communications Magazine 44 (12) (2006) 78–85.
[25] Piero Castoldi, Nicola Samboa, Filippo Cugini, Luca Valcarenghi, QoT-aware lightpath set-up in GMPLS-controlled WDM networks: a survey, Optical Switching and Networking 8 (2011) 275–284.
[26] C. Pinart, N. Sambo, E. Le Rouzic, F. Cugini, P. Castoldi, Probe schemes for quality-of-transmission-aware wavelength provisioning, IEEE/

OSA Journal of Optical Communications and Networking 3 (1) (2011) 87–94.
[27] C. Vijaya Saradhi, S. Subramaniam, Physical layer impairment aware routing (PLIAR) in WDM optical networks: issues and challenges, IEEE Communications Surveys & Tutorials 11 (4) (2009) 109–130.
[28] J. Li, J. Zhang, Y. Zhao, W. Gu, Y. Ji, Impairment-aware backward recursive PCE-based computation algorithm for wavelength switched optical networks, in: Proceedings of 12th IEEE International Conference on the Communication Technology (ICCT), Nanjing, China, 2010.
[29] V. López, B. Huiszoon, J. Fernandez-Palacios, O. Gonzalez de Dios, J. Aracil, Path computation element in telecom networks: recent developments and standardization activities, in: Proceedings of 14th Conference on Optical Network Design and Modeling (ONDM 2010), Kyoto, Japan, 2010.
[30] M. Yannuzzi, E. Marín-Tordera, R. Serral-Gracià, X. Masip-Bruín, González J. Jiménez, D. Verchere, Modeling physical-layer impairments in multi-domain optical networks, in: Proceedings of the 15th International Conference on Optical Network Design and Modeling (ONDM2011), Bologna, Italy, 2011.
[31] X. Yang, B. Ramamurthy, Inter-domain dynamic routing in multilayer optical transport networks, in: Proceedings of the IEEE 2003 Global Communications Conference (Globecom 2003), San Francisco, USA, 2003.
[32] H. Kim, J. Goo Kim, A vertical path computation element based path provisioning scheme in multilayer optical networks, Telecommunication Systems (March) (2011) 1–7. Springer.
[33] S. Gurenben, F. Rambach, Assessment and performance evaluation of PCE-based inter-layer traffic engineering, in: Proceeding of International Conference on Optical Network Design and Modeling (ONDM 2008), Vilanova i la Geltr, Spain, 2008.
[34] F. Cugini, A. Giorgetti, N. Andriolli, F. Paolucci, L. Valcarenghi, P. Castoldi, Multiple path computation element (PCE) cooperation for multi-layer traffic engineering, in: Proceedings of Optical Fiber Communication Conference (OFC 2007), Anaheim, CA, USA, 2007.
[35] M. Miyazawa, S. Kashihara, T. Otani, Optical path computation element interworking with network management system for transparent mesh networks, in: Proceedings of Conference on Optical Fiber communication/National Fiber Optic Engineers Conference OFC/NFOEC (2008), San Diego, CA, USA, 2008.
[36] J.P. Vasseur, J.L. Le Roux, Path Computation Element (PCE) Communication Protocol (PCEP), RFC 5440, 2009.
[37] M. Yannuzzi, X. Masip-Bruín, R. Serral-Gracià, E. Marín-Tordera, A. Sprintson, A. Orda, Maximum coverage at minimum cost for multi-domain IP/MPLS networks, in: Proceedings of IEEE INFOCOM 2008 (Mini-Conference), Phoenix, Arizona, USA, 2008.