Guaranteeing Traffic Survivability and Latency-Awareness in Multilayer Network Design [Invited]

Jose-Juan Pedreno-Manresa, Jose-Luis Izquierdo-Zaragoza, Pablo Pavon-Marino

***Abstract*—The massive adoption of real-time Internet services (e.g., online gaming or video streaming) is threatening the way in which operators and service providers design and operate their networks. New requirements such as maximum latency are becoming de facto indicators to measure network quality, complementing the classical requisites of fault tolerance to increase network availability. However, finding fault-tolerant designs that also limit the end-to-end latency of the IP flows is challenging in multilayer IP-over-WDM networks, since traffic routed over IP links depend on the actual sequences of traversed fibers, and how they change in each failure state. Despite its practical importance, the research on joint optimization of end-to-end latencies and fault tolerance in multilayer optical networks is minimal. In this context, this paper presents a planning algorithm for multilayer IP-over-WDM (with OSPF-ECMP routing in the IP layer), considering both survivability and end-to-end latency-awareness. Three versions of the algorithm are provided, assuming three different recovery mechanisms in the optical layer: 1+1 lightpath protection, unprotected lightpaths, and lightpath restoration. In all the cases, as customary, the optical layer is assumed to react first to the failures, and after that the IP/OSPF layer adapts to the surviving topology. The aim of our algorithm is to design minimum cost networks that guarantee not only 100% survivability under some representative failure states, but also a maximum end-to-end latency for IP flows in any of such failures. According to our results, a careful and holistic design may achieve both objectives simultaneously without, in most cases, noticeable side effects in terms of cost and/or throughput. In its turn, we see that when end-to-end latencies are not considered, designs tend to produce very high latencies in some flows and failure states.**

***Index Terms*—multilayer network design; IP-over-WDM; OSPF; ECMP; latency-awareness; traffic survivability; Net2Plan**

I. Introduction

NOWADAYS the Telecommunication sector is undergoing a major transformation. The emergence of over-the-top (OTT) companies like NetFlix or Google services is challenging traditional operators in many different ways but, most importantly, they are forcing them to evolve their current rigid legacy infrastructures towards more simplified, automated, dynamic and agile architectures, getting rid of static configuration processes [1] and peak-based capacity dimensioning [2]. As a matter of fact, adding more capacity to the network is not enough, and they should guarantee a good quality-of-service/experience (QoS/QoE) for services not under their full control [3].

In this paper we embrace this reality, and contribute with novel design algorithms balancing between costs, revenues and QoS/QoE for IP-over-WDM networks, the industry's standard for operator backbone networks.

The main QoS/QoE metrics for (real-time) multimedia services, like those provided by OTTs, are latency and jitter. The former is the time the information needs to reach end-users from the content generator. The latter is the variance of latency. In this new market scenario, these metrics must coexist with other typical figures in service level agreements (SLAs) like availability levels [4] (e.g. five-nines). Unfortunately, this topic has received limited attention from the community [5].

In IP-over-WDM networks, the end-to-end latency of an IP flow depends not only on the sequence of IP links (lightpaths) traversed, but also on the length of each one of these lightpaths through their physical path. However, IP and optical layers are operated by different personnel and departments, and IP-optical mapping is not a trivial task. Traffic survivability is currently assured through a dual-plane approach, in which two network copies mutually protect each other (1+1 protection), and IP layer is over-provisioned [6]. On the other hand, multilayer restoration has been demonstrated to emulate survivability capabilities at a reduced cost [2]. Anyway, most of existing approaches in the literature does not take control of the latency under failure states.

Several studies have demonstrated the benefits of a joint fault-tolerant design of both optical and IP in multilayer networks [2] in terms of costs; the main contribution of this work consists of improving the design phase to include both fault tolerance and latency-awareness, as an improvement and extension of a previous conference paper [7], with negligible effect in costs.

Three main aspects may be highlighted:

- Although the use of MPLS-TE (Multiprotocol Label Switching - Traffic Engineering) is a raising trend, many networks still use OSPF-ECMP (Open Shortest Path First - Equal-Cost Multi-Path) in the IP layer [8]. This type of routing, while not possessing traffic engineering or individual path control, requires a meticulous and careful design due to weight link tuning [9], and drastic variations in flow routing because of topology changes, i.e. up-to-

Jose-Juan Pedreno-Manresa (e-mail: josej.pedreno@upct.es) and Pablo Pavon-Marino are with the Department of Information and Communication Technologies, Universidad Politécnica de Cartagena, Plaza del Hospital, 1, 30202, Cartagena (Spain).

Jose-Luis Izquierdo-Zaragoza is with Aire Networks, C/ Santiago Ramón y Cajal, 11, 03203, Elche (Spain).

Manuscript received September 28, 2016

down/down-to-up single link events.

- 100% IP traffic survivability is ensured for a set of single-SRG (Shared-Risk Group) failure scenarios. Each SRG represents a reasonable risk we want to be tolerant to, and is composed of the set of links and/or nodes that simultaneously fail when such risk happens. The presented algorithm receives as an input the defined SRGs, which can be arbitrarily defined. In this paper we assume three possible recovery methods for the IP-over-WDM network:
  1) 1+1 optical protection followed by IP restoration, where each IP link is realized by two SRG-disjoint lightpaths.
  2) IP-only restoration, where failed lightpaths are not recovered by the optical layer, and just OSPF reroutes the traffic over the surviving IP layer.
  3) Optical-followed-by-IP restoration (or multilayer restoration), which involves cooperation between the two layers. Here, the optical layer attempts first to restore failing lightpaths. After that, the IP/OSPF layer reroutes the IP traffic in the surviving lightpaths.
- Lastly, the algorithm ensures a maximum end-to-end latency for all IP traffic even *under any failure state*.

The algorithm will be thoroughly tested with different reference topologies and traffic matrices using the open-source network planner Net2Plan [10], [11]. The case study will focus on analyzing the network cost and throughput for different strategies, and the impact that imposing latency constraints has on those metrics. Results will illustrate the relevance of applying latency constrains to network design to avoid significantly high end-to-end latency under certain failure states.

The rest of the paper is organized as follows: in Section II we review previous works in multilayer network planning algorithms regarding their survivability and latency requirements. Section III contains a description of the proposed multilayer planning algorithm, considering three different recovery methods and assuring a maximum end-to-end latency. In Section IV, we report and discuss the results obtained in a series of test from our case study. Finally, Section V concludes the paper.

II. Related Work

Traffic survivability is of paramount importance in network design, and has been thoroughly discussed in multilayer network design literature [12]. As aforementioned, in the light of emergent new technologies, services and paradigms, there is still room for research and improvement. For example, joint consideration of traffic survivability and energy efficiency was analyzed in the literature some years ago [13].

This section makes a brief review of previous works about survivability and latency-aware design that motivated this paper.

A. Survivability-Aware Design in Multilayer Networks

One of the classical scenarios in IP/OSPF-over-WDM networks is the one where no recovery scheme is used in the optical layer, so in case of failure, it is the duty of the IP layer to reroute disrupted traffic on the surviving topology. In this case, enough capacity (extra lightpaths) must be provisioned beforehand considering all possible failure states. A brief survey about the different dimensioning techniques in this scenario is presented in [8]. The different discussed approaches differ among themselves on the level of integration of the different components of a multilayer network design problem (e.g. traffic routing, design and dimensioning of the IP topology). The authors conclude that considering the previous mentioned factors altogether helps to maximize the benefits of multilayer networking.

The application of multilayer restoration as a recovery scheme is discussed in [14]. In the proposed strategy, recovery begins in the optical layer (reallocating failing lightpaths) and, then, escalates to the IP layer (rerouting still disrupted traffic). Thus, blocked traffic (i.e. offered traffic between two nodes that cannot be allocated) that could not be recovered in the optical layer can still be it in the upper layer. Then, multilayer restoration is compared to traditional IP-only restoration, showing that the former is a cost-efficient alternative in terms of optical transponders. Coordination between layers is mandatory to ensure a correct process of restoration, avoiding that both layers take recovery actions at the same time against the same failure. To do so, hold-off timers are defined to guarantee the IP layer does not trigger rerouting until a certain time has passed, so the optical layer has finished its reaction against this failure.

On the other hand, authors in [15] provide another perspective by comparing multilayer restoration to dedicated optical layer protection (1+1 optical protection). Again, the former recovery scheme shows significant cost savings.

B. Latency-Aware Design in Multilayer Networks

End-to-end latency is becoming an important QoS/QoE in SLA contracts. Despite this fact, in multilayer network planning, it is usually addressed in a best-effort manner. Typical approaches involve routing traffic over the shortest paths across all layers and then dimension IP links accordingly [8]. On the optical layer, lightpath routing is often performed using shortest path algorithms [16] like the well-known Dijkstra's algorithm [17] or related variants, such as Suurballe's algorithm [18] for node/link disjoint dedicated path protection [19].

In fact, even in single-layer IP networks, it is very difficult and challenging to enforce a maximum end-to-end latency value, since OSPF-ECMP routing requires a careful tuning of link weights to achieve TE and/or SLA objectives [9]. Unfortunately, this type of problem is not suitable to be solved using Integer Linear Programming (ILP) techniques; besides, path-based constraints like end-to-end latency cannot be easily introduced into ILPs for hop-by-hop routing schemes like OSPF-ECMP. In addition, heuristic-based approaches for weight tuning may not be able to satisfy QoS/SLA and/or TE requirements unless objective functions consider them explicitly [20]. However, it is important to remark that network operators prefer assigning naive *static* link weight strategies

(e.g. hop-based or distance based approaches) rather than perform weight tuning according to network conditions [8].

Interestingly, research on latency-aware design has largely focused on MPLS routing for the IP layer, since paths can be explicitly defined for each IP demand according to TE and SLA objectives [21]. As an example, authors present in [22] a network design based on explicit routing to ensure equalization of end-to-end latency between different IP demands, while keeping a maximum end-to-end latency value. Unfortunately, they focus only on single-layer IP networks.

C. Joint Fault Tolerant and Latency-Aware in Multilayer Networks

Regarding to the scope of this paper, there are few relevant works in the literature considering both aspects.

Even for single-layer IP networks, few works can be found for link weight tuning in IP/OSPF networks considering survivability and latency-awareness. Authors in [23] present an heuristic algorithm to compute the optimal link weights to ensure network survivability against a predefined set of failures. Although latency constraints are not defined, authors give guidelines to include them.

Regarding multilayer networks, the necessity of considering survivability and latency constraints in joint design has been discussed and motivated in [5]. One of the few works that proposed joint design considering both aspects is presented in [11]. Authors propose a multilayer algorithm which guarantees 100% survivability and a maximum end-to-end latency, applying OSPF-ECMP routing in the IP layer and 1+1 optical protection in the optical layer (using two disjoint lightpaths for each IP link). A modified version of the well-known IGP-WO (Interior Gateway Protocol - Weight Optimizer) algorithm [9] is used to find the link weights that minimize the network congestion. The objective function is tuned (as proposed in [22]) to penalize solutions that do not meet the latency criteria.

It is worth noting the lack of proposals for hybrid OSPF-MPLS routing in multilayer network which deal with this problem. These hybrid routing schemes received some attention in the past, for both protected [24] and unprotected [25] design in single-layer IP networks. Basically, two different IP routing schemes are applied to different traffic profiles: high-priority traffic (with strict QoS/SLA requirements) is routed using fine-grained MPLS paths, whereas the low-priority traffic is routed using OSPF. Adapting this technique to multilayer networks would imply MPLS routing at the IP layer and 1+1 optical protection at the optical layer for high-priority traffic, while the remaining traffic is routed through OSPF.

Finally, this paper extends and improves preliminary work presented in [7]. Essentially, the core of the algorithm has been completely rewritten, adjusting several heuristic decisions with the objective of achieving a lower cost and increase the overall throughput. In addition, more exhaustive and illustrative tests are provided in order to compare network designs with and without latency limitation, and to thoroughly explore the impact on cost and achievable throughput.

As a last point, while applying three different recovery schemes, OSPF link weights are left untouched during network operation. Therefore, network operators can apply their weight criteria in a much broader set of situations.

To conclude this section, Table I collects and briefly summarizes the contributions of previously discussed works in the areas of interest of this paper.

TABLE I
CLASSIFICATION OF MULTILAYER NETWORK DESIGN ALGORITHMS

	Survivability-Aware	Survivability-Unaware
Latency-Aware	[7] [11] [23]	[21] [22]
Latency-Unaware	[8] [14] [15] [16]	N/A

III. ALGORITHM DESCRIPTION

In this section, we describe our proposal. The aim of our multilayer algorithm is to create minimum cost IP-over-WDM designs where the 100% of the IP traffic survives to a set of predefined failure states, whereas guaranteeing the end-to-end latency at the IP level will not exceed a given threshold in any possible network state, either faulty or failure-free. The algorithm considers three different recovery schemes.

The input data of our algorithm are:

- IP-over-WDM topology ($G(N, E)$), composed of a set of multilayer-capable nodes ($n \in N$) interconnected via unidirectional fibers ($e \in E$)
- A set of arbitrary-defined SRGs ($r \in R$), representing the set of possible failures. Without loss of generality, we assume the network will remain connected after any single SRG failure (i.e. a physical path can be found between each node pair), so that guaranteeing 100% survivability is always possible. The set of network states ($s \in S$) is composted of the failure-free state and all single SRG failures.
- A set of end-to-end IP demands ($d \in D$), representing the end-to-end traffic demands between node pairs.
- Maximum end-to-end latency value (L_{max}) to be met by any IP demand ($d \in D$) under any network state ($s \in S$). Only propagation delays at the traversed lightpaths are considered for computing the end-to-end latencies.

In order to analyze traffic survivability and latency constraints, our algorithm considers the reaction of the network through three different recovery mechanisms: 1+1 optical protection followed by IP restoration, optical-followed-by-IP restoration (also known as multilayer restoration [26]) and IP-only restoration. In this way, we are able to dimension the IP layer according to the expected behavior of the network.

As mentioned before, the routing in the IP layer is governed by OSPF-ECMP according to the current standards, having all links a weight equal to one. At the optical layer, it is assumed WDM technology, with no wavelength conversion and no need for intermediate regeneration.

A. Implementation

Multilayer network design variants are $\mathcal{NP}$-complete problems, and we resort to a heuristic algorithm in our algorithm

design, based on the Greedy Randomized Adaptive Search (GRASP) [27] scheme. GRASP is a well-known iterative meta-heuristic, composed of two differentiated phases per iteration: (i) construction and (ii) local search. In the first phase, a feasible solution is constructed using a greedy algorithm. In our case, we consider as a feasible solution a network design which carries all the offered traffic, ensures 100% survivability in case of single SRG failure and meets a maximum end-to-end latency in all states. In case a solution cannot be found (i.e. one or more lightpaths cannot be allocated), the algorithm starts a new iteration. Once a solution is found, the second phase tries to improve the objective function using a local search algorithm, in this case reducing the cost by minimizing the total number of transponders. After finishing all iterations, the best solution found so far is returned.

The algorithm is composed of several hierarchical modules (see Fig. 1). Each module is tasked to take pre-defined actions and delegating into other modules events outside its scope. The main module, contains the aforementioned GRASP scheme. The IP-over-WDM module is used to apply RWA (Routing and Wavelength Assignment) and recovery mechanisms as it would be in a real-world scenario. Adding or removing lightpaths is done by the WDM module, and the IP module is tasked with recalculating OSPF-ECMP rules at the IP layer.

For example, the main module sends an event to the IP-over-WDM module (e.g. 'Add-lightpath' or 'SRG-failed'), which reacts by taking some actions or sending new events to the WDM and/or IP module depending of the type of event received. This procedure allows us to generate offline planning algorithms by reusing code from existing online algorithms, for the purpose of dynamic lightpath allocation, setting up recovery schemes, and so on.

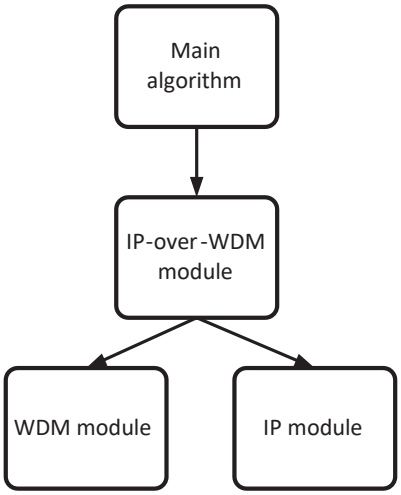

Fig. 1. Algorithm Modules Hierarchy

The construction phase starts with the input data described in Section III, and its pseudo-code is presented in Fig. 2. Initially, the network is empty, and no traffic is carried. In each iteration of the main loop, one lightpath is added, and the traffic is routed as an IP-over-WDM network makes. The `evalConstraints()` method, receives a design, and evaluates whether the fault-tolerance and end-to-end latency constraints are met. It returns the set of IP demands with non-carried traffic (BT_D), for which no IP path exists, the set of IP links with oversubscribed traffic ($OT_{E_{IP}}$), and the set of IP demands which violate the maximum latency constraint ($maxL_D$), storing the worst-case among all states. The construction phase ends by leaving the main loop, when (i) no constraint is violated, that is, when BT_D, $OT_{E_{IP}}$ and $maxL_D$ become empty, or (ii) no additional lightpath can be setup leading to a new iteration of the main algorithm.

Algorithm 1 Construction phase

Require: $G(V, E)$, D, S, L_{max}
  [BT_D, $OT_{E_{IP}}$, $maxL_D$] = `evalConstraints()`
  while $BT_D \neq \emptyset$ or $OT_{E_{IP}} \neq \emptyset$ or $maxL_D \neq \emptyset$ **do**
    if $BT_D \neq \emptyset$ **then**
      select randomly weighted an ingress-egress pair of nodes with blocked traffic
    else if $OT_{E_{IP}} \neq \emptyset$ and $random\_number < \alpha$ **then**
      select randomly weighted an ingress-egress pair of nodes with oversubscribed IP links between them.
    else
      select randomly weighted an ingress-egress pair of nodes with non-allowed latency between them
    end if
    establish a new IP link (lightpath) between the selected node pair
    [BT_D, $OT_{E_{IP}}$, $maxL_D$] = `evalConstraints()`
  end while

Fig. 2. Pseudo-Code for the Construction Phase

Digging into the `evalConstraints()` method, we iterate over all possible network states ($s \in S$) to check the fault tolerance and latency constraints in all of them. To do so, since every failure state is modeled as an SRG, an 'SRG-failed' event is sent to the IP-over-WDM module which reproduces the reaction of the IP-over-WDM network when transiting from a non-failure state, to the failure state s. This depends on the recovery mechanism considered, as described in Section III-B. Then, the resulting state is checked, some internal metrics (those defined in the previous paragraph) are computed and the network is reverted to normal operation (no failure).

After execution of `evalConstraints()`, if at least one constraint is violated (so the design is not still valid), we proceed as follows. First, we check the presence of pair of nodes with blocked traffic (BT_D), selecting among those a pair of ingress-egress nodes using a weighted random method, where each weight value is the amount of blocked traffic. In case there is no blocked traffic, if some link is oversubscribed we pick a random number and compare it to an input parameter α to determine if a new lightpath must be setup between oversubscribed links. The rationale behind this is to modulate the probability of adding too many lightpaths. If we are allowed to add new lightpaths, then randomly select a node-pair from those in $OT_{E_{IP}}$, using the oversubscribed traffic as weight. Finally, if none of the previous options was selected, we check whether exists any IP link for any demand ($maxL_D$) exceeding the maximum end-to-end latency value (see Section

III-C). Again, a weighted random selection is performed using the excess latency (in milliseconds) as weight.

Once a node pair has been selected (through any of the conditions mentioned above), a new IP link (lightpath) is established between the ingress and the egress node. To allocate a new lightpath, our RWA includes a fixed-alternate routing strategy along with a first-fit wavelength allocation. In other words, we have a precomputed set of candidate paths between each node pair and try to find the shortest path selecting the common free wavelength with the lower index. Due to the RWA constraints, this part of the algorithm might be unable to allocate a lightpath; this inability to provide a feasible design stops the construction phase.

Finally, the constraints are computed again, and the whole process is repeated until a feasible solution is found (with no constraint violation) or no more lightpaths can be established (invalid solution). If a feasible solution can be found, then the local search stage is started.

The purpose of the local search phase is to improve the cost of a feasible solution coming from the first phase, removing as many IP links as possible still satisfying the survivability and latency requisites. Fig. 3 shows the corresponding pseudo-code. First, all IP links are ordered in terms of the spare capacity in descending order (E_{IP}). Then, we iterate over them, assuming in each iteration the selected IP link is removed, and evaluating the results of such a removal with `evalConstraints()`. In case no constraint is violated in any of the failure states, the IP link is removed. This process is repeated until no more IP links can be eliminated.

Algorithm 2 Local search phase

Require: A feasible solution from Construction phase
  repeat
    order all IP links by decreasing order of spare capacity (denoted as E_{IP})
    for all IP link $e_{IP} \in E_{IP}$ **do**
      mark e_{IP} as removed
      [BT_D, $OT_{E_{IP}}$, $maxL_D$] = `evalConstraints()`
      if $BT_D = \emptyset$ and $OT_{E_{IP}} = \emptyset$ and $maxL_D = \emptyset$ **then**
        remove IP link e_{IP}
        exit **for** loop
      else
        unmark e_{IP}
      end if
    end for
  until no IP link can be removed

Fig. 3. Pseudo-Code for the Local Search Phase

B. IP-over-WDM Module and Recovery Schemes

We developed the IP-over-WDM module to support the three recovery schemes considered, using a bottom-up multi-layer approach [16] (the same applied in [14]); that is, when an event is received, the main module first invokes actions from the WDM module to compute the reactions to the failure of the optical layer, and then calls the IP module to obtain the reaction of the IP layer to the surviving topology. The optical module actions depend on the selected recovery scheme, while the scheme computing the OSPF-ECMP reactions at the IP module is common for all recovery mechanisms (i.e rerouting according to the surviving IP topology).

As for the WDM module, when using 1+1 optical protection followed by IP restoration, each new 'Add-lightpath' event is treated realizing each IP link with two SRG-disjoint lightpaths. Therefore, in case of an 'SRG-failed' event, no actions must be taken and traffic survivability is guaranteed as lightpaths were created as SRG-disjoint.

For multilayer restoration, the WDM module is in charge of applying RWA strategies to allocate new lightpaths, and in case of failure event, to reroute as much failing lightpaths as possible over the surviving topology before the IP module recomputes the OSPF-ECMP rules.

Finally, for IP-only restoration, the WDM module is only in charge of allocating new lightpaths and no action is taken at the optical layer when a failure occurs, since failing lightpaths are not restored.

C. Latency Computation

A thorough computation of end-to-end latency metric would involve taking into account the propagation delay, queuing delay, transmission delay, and packet processing times. Assuming the network is properly dimensioned, congestion episodes will be minimal and queuing delay will be minimized. Moreover, supra-gibabit/second line rates and packet processing at wire-speed leave the propagation delay as the main contribution to the latency in backbone networks [28]. Therefore, for the sake of simplicity, we will only consider propagation delay as the sole source of latency.

For this study, since IP/OSPF-ECMP is based on hop-by-hop routing, there is a need to obtain explicit end-to-end paths in order to compute per-flow propagation delays. Given an IP network, a set of IP demands, and IP link weights, we apply path reconstruction method inspired by Edmonds-Karp algorithm [29] to compute end-to-end paths from ECMP forwarding rules. A more detailed explanation can be found in [30]. This computation is extensively used in `evalConstraints()` to retrieve the worst-case end-to-end delay among all paths carrying traffic for each IP demand and each failure scenario.

IV. Case Study

In this section, we report illustrative results collected from a series of tests. We aim to analyze for different network scenarios, different trade-offs appearing respecting to the overall throughput (or maximum carried traffic before fiber resource exhaustion), network cost and maximum end-to-end latency depending on the selected recovery schemes (described in Section III-B).

These tests were performed using the offline network design tool from Net2Plan [11]. This tool allows users to design and dimension networks assuming some static information, as well as obtaining statistics and other significant data. The inputs to these tests are the optical physical topology, an IP traffic matrix and the planning algorithm with its parameters.

The algorithm is developed in Java, implementing public and well documented interfaces. We remark that for the purpose of inspection and validation, Net2Plan and the source code of the algorithm are publicly available on the websites [10], [31].

A. Testing Scenario

We used three different well-known IP-over-WDM networks topologies composed of (unidirectional) fiber links and network sites including multilayer equipment: NSFNet [32], Internet2 [33] and Atlanta [34]. The IP traffic for each network was generated using the population-distance model described in [35]. We assume each node is equipped with IP routers and ROADM equipment with 100 Gbps transponders, and 40 WDM channels per fiber. Lastly, we consider the propagation speed of light in the fiber to be 200000 km/s and no need for signal regeneration between node pairs. A brief summary of the size of each network can be seen in Table II.

TABLE II
NETWORK TOPOLOGIES

Network	Nodes	Links	Average node degree
NSFNet	14	42	3
Internet2	9	26	2.889
Atlanta	15	44	2.933

To standardize all tests, fiber lengths in all networks have been appropriately scaled to normalize the network diameter to 5000 km in all cases, which represents an end-to-end delay of 25 ms. We note that the network diameter is the longest (in this case, measured in kilometers) among all-pairs shortest paths.

We run the algorithm in each topology, (i) for different scaled versions of its seminal IP traffic matrix, with total offered traffic ranging from 500 Gbps to 16 Tbps in steps of 500 Gbps, and (ii) different end-to-end latency limits: 50 ms, 62.5 ms, 75 ms, and unbounded (no limit).

One SRG is defined for each bidirectional fiber pair between each node pair. In this form, designs should be tolerant to single duct cuts, where the duct cut breaks simultaneously the two fibers in opposite directions in the duct [36]. This kind of failure may be caused by natural disasters, human error (civil engineering activities), or even represent the case of a programmed maintenance (e.g. upgrade of optical line amplifiers). Note that the algorithm can easily accommodate multiple duct cuts, or other particular failure situations, by defining SRGs accordingly.

Once all the executions have finished we collect the total necessary number of transponders to establish the planned design for each of the recovery mechanisms. As occurs in practical deployments [37], transponders are supposed to be bidirectional, and transmission and reception lightpaths can be tuned at different wavelengths and follow different routes from the same end nodes. According to this, the number of transponders used in a node is the maximum between all the number of incoming and outgoing lightpaths to the node. As is customary in the literature, the number of transponders, which is tightly related to the number of necessary IP ports, is considered a good approximation of the overall network cost [38]. Finally, we collect the worst case of end-to-end latency for each IP traffic demand under any of single SRG failures and the total throughput achieved with each recovery scheme and latency limitation.

B. Results

Tables III, IV and V show the overall throughput achieved for each of the networks under the three recovery schemes: 1+1 optical protection (1+1), IP-only restoration (IP-R) and optical-followed-by-IP restoration (Op-IP-R); and the different imposed latency limitations. Network throughput is computed as the sum of the traffic of the maximum IP traffic matrix for which a feasible survivable and latency aware design was found.

It is interesting to remark that a maximum allowed end-to-end latency of three times the normalized network diameter (75 ms) has no effect on the throughput respect to no limitation, no matter the selected recovery method. Even a limit of 62.5 ms (two times and a half the network diameter) of end-to-end latency has no noticeable difference, except on the case of 1+1 optical protection. The most consistent throughput is obtained using IP-only restoration, this may be due to the fact that in case of failure, no new lightpaths have to be allocated, removing the possibility of not having a sequence of links that meets the latency criteria.

TABLE III
OVERALL THROUGHPUT ON NSFNET (IN TBPS)

Latency limit	1+1	IP-R	Op-IP-R
No limit	9.5	11.5	16.0
75.0 ms	9.0	11.5	16.0
62.5 ms	6.0	11.5	15.5
50.0 ms	N/A	7.0	6.0

TABLE IV
OVERALL THROUGHPUT ON INTERNET2 (IN TBPS)

Latency limit	1+1	IP-R	Op-IP-R
No limit	8.0	7.0	10.0
75.0 ms	8.0	7.0	10.0
62.5 ms	8.0	7.0	10.0
50.0 ms	8.0	7.0	9.5

Regarding the cost in terms of the necessary number of transponders, Figs. 4, 5 and 6 show a comparison for each network between a design with no latency limit and a maximum limit of 62.5 ms for each of the three recovery schemes. Each value represents the total cost for a valid design which ensures 100% survivability.

Comparing the three recovery methods we can draw several interesting conclusions. First, there is a consistent ranking from highest to lowest cost for any given offered traffic in terms of transponder requirements: 1+1 optical protection,

TABLE V
OVERALL THROUGHPUT ON ATLANTA (IN TBPS)

Latency limit	1+1	IP-R	Op-IP-R
No limit	5.5	8.5	9.5
75.0 ms	5.5	8.5	9.5
62.5 ms	3.0	8.5	9.5
50.0 ms	N/A	8.5	6.0

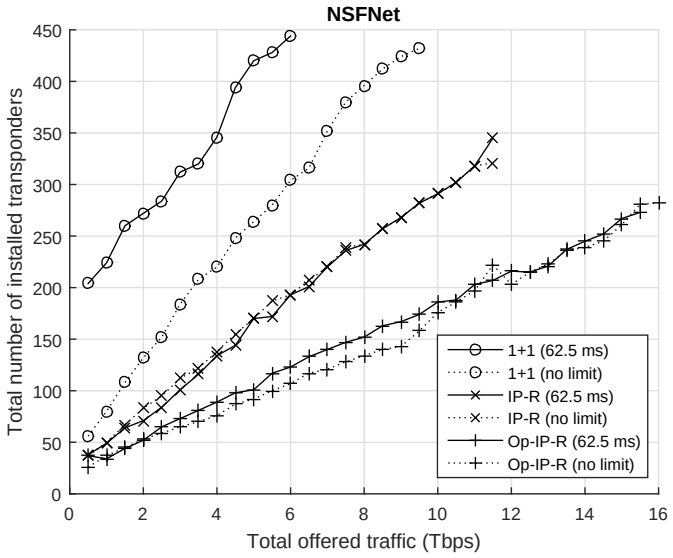

Fig. 4. Total Number of Transponders for NSFNet (No Latency Limit vs Maximum Limit of 62.5 ms)

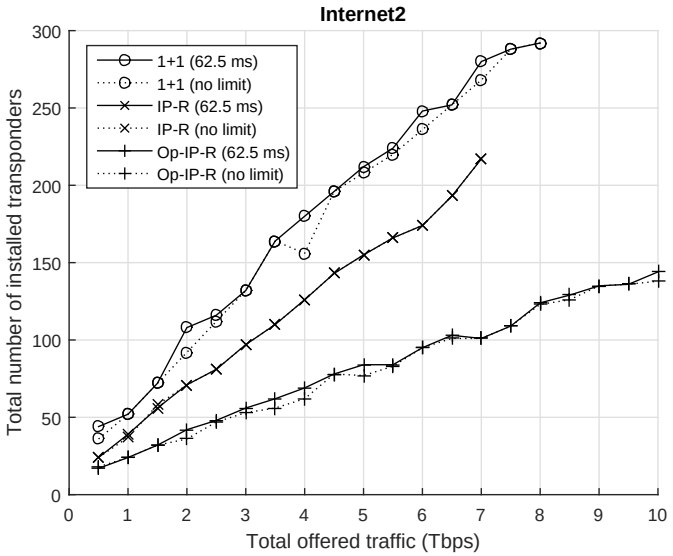

Fig. 5. Total Number of Transponders for Internet2 (No Latency Limit vs Maximum Limit of 62.5 ms)

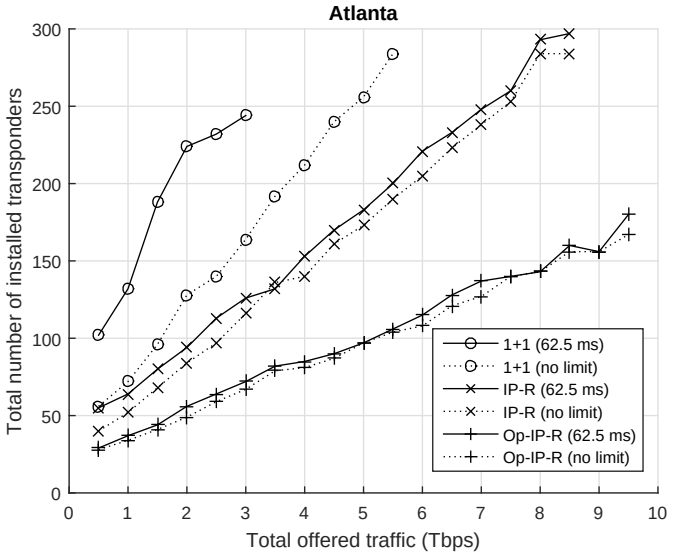

Fig. 6. Total Number of Transponders for Atlanta (No Latency Limit vs Maximum Limit of 62.5 ms)

followed by IP-only restoration and optical-followed-by-IP restoration. This behavior is expected, 1+1 optical protection realizes each IP link has two different SRG-disjoint lightpaths, and therefore requires two pair of transponders. In the case of IP-only restoration, additional lightpaths must be allocated before network operation to ensure no traffic is disrupted in case of failure, causing an network over-dimensioning. Finally, optical-followed-by-IP restoration is able to reroute failing lightpaths over the surviving physical topology and therefore some failures are unnoticed to the IP layer, being the trade-off a moderate over-dimensioning.

Second, from the point of view of a fixed number of transponders, the most efficient recovery scheme is optical-followed-by-IP restoration which achieves the maximum throughput compared to the two other alternatives.

Lastly, we remark that compared to our previous work [7], this algorithm achieves better throughput in all recovery schemes with a noticeable lower cost (20% average less transponders), thanks to the improvement of the construction and local phases of the algorithm. Compared to the previous version, the new construction phase prioritizes the allocation of new lightpaths taking into account the amount of blocked traffic, link oversubscribe and deviation from the maximum allowed end-to-end latency while maintaining certain randomness. A more aggressive approach in the local phase allows a decrease in transponders respect to our previous work.

Comparing for each recovery scheme (i) the lack of latency constraining to (ii) an imposed limit equivalent to 2.5 times the propagation delay of the network diameter, we can see some interesting findings. First of all the most observable difference is found in 1+1 optical protection. Using this kind of recovery while ensuring a maximum end-to-end latency at the IP layer leads to a notable increase of the number of transponders as well as a much reduced total throughput (being the exception the Internet2 network). As for optical-followed-by-IP restoration and IP-only restoration there is almost no increase on the number of transponders or decrease in the overall throughput (although in some particular cases it may seem a lower cost when imposing a latency limit, this may be caused to the random nature of the algorithm). It is safe to say that guaranteeing a maximum end-to-end latency on IP traffic without a perceptible increase in cost is always desirable.

Figs. 7, 8 and 9 provides us with a further detail on the end-to-end latency behaviors in the network. They illustrate the histograms for worst-case (including failure states) latency distribution for each network for a given offered traffic of 3 Tbps. Each figure contains two sets of histograms, the upper one represents the worst-case end-to-end latency among all failure states for each recovery scheme, assuming no imposed limit to latency. The lower set contains the same information, but limiting the maximum end-to-end latency to 62.5 ms. A vertical dashed lined in each histogram highlights the 62.5 ms limit. We can easily observe that in the latency-aware scenario, the algorithm succeeds in making every flow meet the end-to-end latency constraint. In the case of unrestricted latency design, although a good proportion of IP flows meets the constraint criteria, many others exceed the maximum given latency, even by a large extent. This suggests that not considering latency limitations in the network design can easily produce large maximum latency violations.

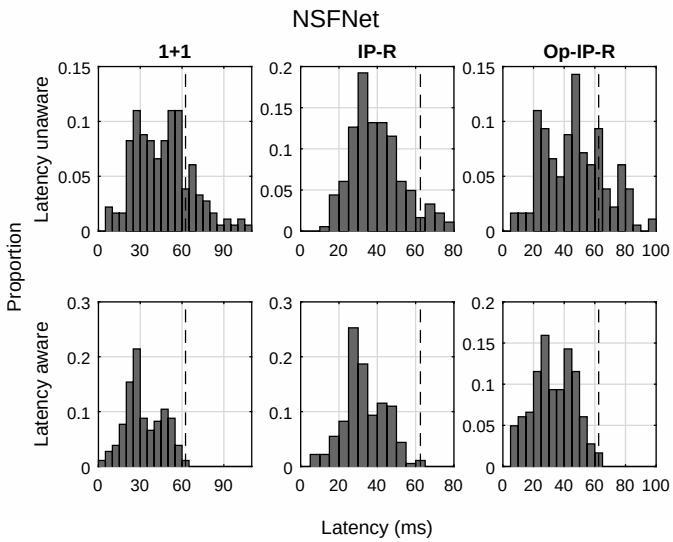

Fig. 7. Worst-Case End-to-End Latency Distribution for NSFNet (No Latency Limit vs Maximum Limit of 62.5 ms)

Tables VI, VII and VIII show a more detailed information about the proportion of IP demands and offered traffic exceeding maximum latency in case of an unrestricted design. We observe that without considering latency-aware limitations, a significant proportion of IP traffic will suffer from a high latency, either during normal operation or failure state. For instance, this impacts intensely to the optical-followed by IP restoration case. This case has been proven as the best solution both from lower cost and overall throughput perspective. However, its design should be specially careful, since a network design not considering latency limitations may result in an unacceptable proportion of traffic exceeding the imposed maximum latency, as seen on NSFNet and Internet2 networks. Again, this phenomenon does not occur when applying latency-aware design thus supporting and validating the correct behavior of our proposed joint algorithm.

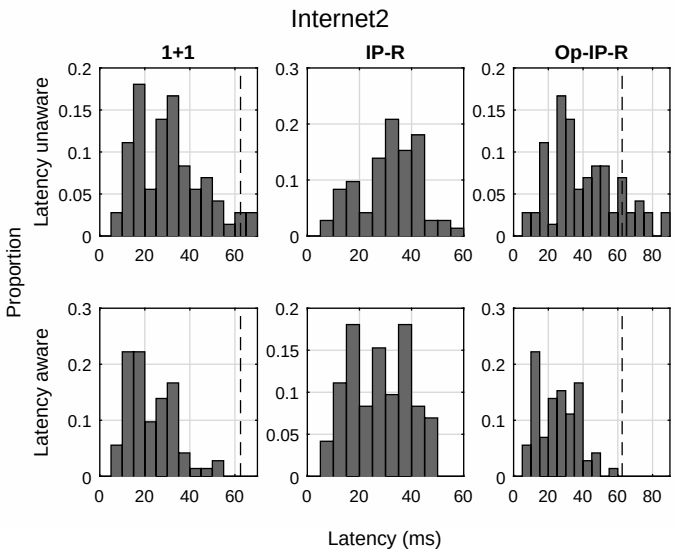

Fig. 8. Worst-Case End-to-End Latency Distribution for Internet2 (No Latency Limit vs Maximum Limit of 62.5 ms)

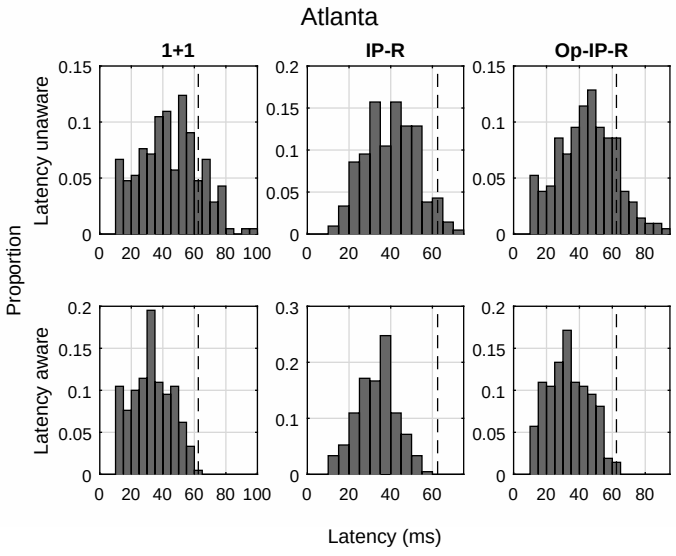

Fig. 9. Worst-Case End-to-End Latency Distribution for Atlanta (No Latency Limit vs Maximum Limit of 62.5 ms)

TABLE VI
PROPORTION OF IP DEMANDS AND OFFERED TRAFFIC VIOLATING LATENCY CONSTRAINTS ON NSFNET (LATENCY-UNAWARE)

Recovery scheme	% IP Demands	% O.T.	O.T. (in Gbps)
1+1	19.23	19.46	584
IP-R	7.14	6.1	183
Op-IP-R	23.62	23.2	696

*O.T. = Offered Traffic

C. Discussion and Further Work

Results shown in this paper clearly validate optical-followed-by-IP restoration (or multilayer restoration) as the winning recovery scheme both in terms of cost and overall throughput. These results for the joint fault-tolerant and latency-aware scenario addressed in this paper, are consistent to other previous results in less restricted scenarios. Still, 1+1-protected and unprotected lightpaths are still dominant

TABLE VII
PROPORTION OF IP DEMANDS AND OFFERED TRAFFIC VIOLATING LATENCY CONSTRAINTS ON INTERNET2 (LATENCY-UNAWARE)

Recovery scheme	% IP Demands	% O.T.	O.T. (in Gbps)
1+1	4.16	6.13	184
IP-R	0	0	0
Op-IP-R	13.88	13.33	400

*O.T. = Offered Traffic

TABLE VIII
PROPORTION OF IP DEMANDS AND OFFERED TRAFFIC VIOLATING LATENCY CONSTRAINTS ON ATLANTA (LATENCY-UNAWARE)

Recovery scheme	% IP Demands	% O.T.	O.T. (in Gbps)
1+1	18.75	9.5	285
IP-R	4.76	1.96	59
Op-IP-R	14.7	5.22	158

*O.T. = Offered Traffic

in many telco networks, given the difficulty to perform and agile reconfiguration at the optical layer, required by lightpath restoration. In addition, even when possible, lightpath restoration times can be slow (e.g. tens of seconds to minutes) in current optical infrastructures, due to the need of channel equalization processes [2]. Moreover, OSPF reconvergence is in the order of dozens of seconds [39], so recovery time in case of optical-followed-by-IP restoration is at least as bad as IP-only restoration. On the other hand, 1+1 optical protection realizes each IP link as two disjoint lightpaths, so neither setup time nor reconvergence time is needed in case of single failure, being this method the fastest of the three. A good compromise between cost/throughput and recovery time would be the application of IP fast re-routing techniques (IPFRR) which guarantee that OSPF reactions for rerouting IP traffic occur at a sub-second time. Therefore, further work includes the application of the aforementioned techniques to maximize IPFRR coverage [40] while maintaining end-to-end latency constraints for IP traffic.

As a further work, it is of practical importance investigating algorithms that impose the latency restrictions selectively so some IP demands have different limits than others, prioritizing that way certain flows in terms of latency. In this scenario, the bottom-up restoration may not work as expected, and top-down approaches may become worth to analyze, provided that multilayer negotiation and coordination is not considered in the short-term. In fact, authors in [5] propose the application of IP-only restoration for sub-second recovery of high-priority services, followed by a combination of IP-optical mechanisms for the rest of the traffic.

V. CONCLUSIONS

In this paper we present a planning algorithm for IP-over-WDM networks which jointly guarantees fault-tolerance to a selected set of failures, and a maximum end-to-end latency for IP traffic taking into account three different recovery schemes. This algorithm, improves both costs (in terms of optical transponders) and overall throughput from our previous work [7]. A set of extensive results have been provided, that validate our proposal and illustrate the possibility of creating joint designs that ensure survivability while avoiding IP flows to suffer excessive end-to-end latencies.

In our results, we have also seen how with a careful multilayer design, it is possible to find fault-tolerant and latency-aware designs, with a small fraction of extra cost respect to fault-tolerant designs where end-to-end latency limits are not met.

ACKNOWLEDGMENT

This work was partially supported by the FPU fellowship program of the Spanish Ministry of Education, Culture and Sports (ref. FPU14/04227), by the Spanish project grants TEC2014-53071-C3-1-P (ONOFRE) and TEC2015-71932-REDT (ELASTIC), and by the Institut Valencià de Competitivitat Empresarial and the European Regional Development Fund through the project IFITDA/2015/19 (Net^{2}Evolution).

REFERENCES

[1] L. Velasco, A. Castro, D. King, O. Gerstel, R. Casellas, and V. Lopez, "In-Operation Network Planning," *IEEE Communications Magazine*, vol. 52, no. 1, pp. 52–60, Jan. 2014.
[2] O. Gerstel *et al.*, "Multi-Layer Capacity Planning for IP-Optical Networks," *IEEE Communications Magazine*, vol. 52, no. 1, pp. 44–51, Jan. 2014.
[3] T. Janevski, "QoS/QoE frameworks for converged services and applications," in *Proceedings of the Regional Workshop for Europe "New Issues in Quality of Service Measuring and Monitoring"*, Bologna (Italy), Nov. 2015.
[4] J. Rak, *Resilient Routing in Communication Networks*, 1st ed., ser. Computer Communications and Networks. Springer, Nov. 2015.
[5] O. Gerstel, C. Filsfils, and W. Wakim, "IP-Optical Interaction during Traffic Restoration," in *Proceedings of the Optical Fiber Communication Conference and Exposition and the National Fiber Optic Engineers Conference 2013 (OFC/NFOEC 2013)*, Anaheim, CA (United States), Mar. 2013.
[6] J.-P. Fernandez-Palacios *et al.*, "IP and Optical Convergence: Use Cases and Technical Requirements," Telefónica I+D, AXTEL Mexico, Bouygues Telecom, BT, China Unicom, Colt, Deutsche Telekom, KDDI, Korea Telecom, Orange and Telecom Italia, White Paper, Jan. 2014.
[7] J.-J. Pedreno-Manresa, J.-L. Izquierdo-Zaragoza, and P. Pavon-Marino, "Joint Fault Tolerant and Latency-Aware Design of Multilayer Optical Networks," in *Proceedings of the 20th International Conference on Optical Network Design and Modeling (ONDM 2016)*, Cartagena (Spain), May 2016.
[8] E. Palkopoulou, O. Gerstel, I. Stiakogiannakis, T. Telkamp, V. Lopez, and I. Tomkos, "Impact of IP Layer Routing Policy on Multilayer Design [Invited]," *IEEE/OSA Journal of Optical Communications and Networking*, vol. 7, no. 3, pp. A396–A402, Mar. 2015.
[9] B. Fortz and M. Thorup, "Internet Traffic Engineering by Optimizing OSPF Weights," in *Proceedings of the 19th Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM 2000)*, Tel Aviv (Israel), Mar. 2000.
[10] "Net2Plan – The open-source network planner," [Last accessed: January 2017]. [Online]. Available: http://www.net2plan.com
[11] P. Pavon-Marino and J.-L. Izquierdo-Zaragoza, "Net2Plan: An Open Source Network Planning Tool for Bridging the Gap between Academia and Industry," *IEEE Network*, vol. 29, no. 5, pp. 90–96, Sep.-Oct. 2015.
[12] P. Demeester *et al.*, "Resilience in Multilayer Networks," *IEEE Communications Magazine*, vol. 37, no. 8, pp. 70–76, Aug. 1999.
[13] A. Jirattigalachote, C. Cavdar, P. Monti, L. Wosinska, and A. Tzanakaki, "Dynamic provisioning strategies for energy efficient WDM networks with dedicated path protection," *Optical Switching and Networking*, vol. 8, no. 3, pp. 201–213, Jul. 2011.
[14] J.-L. Izquierdo-Zaragoza and P. Pavon-Marino, "Assessing IP vs optical restoration using the open-source Net2Plan tool," in *Proceedings of the 16th International Telecommunications Network Strategy and Planning Symposium (NETWORKS 2014)*, Funchal (Portugal), Sep. 2014.

[15] A. Mayoral, V. Lopez, O. Gerstel, E. Palkopoulou, O. Gonzalez de Dios, and J.-P. Fernandez-Palacios, "Minimizing Resource Protection in IP Over WDM Networks: Multi-layer Shared Backup Router [Invited]," *IEEE/OSA Journal of Optical Communications and Networking*, vol. 7, no. 3, pp. A440–A446, Mar. 2015.

[16] J.-P. Vasseur, M. Pickavet, and P. Demeester, *Network Recovery: Protection and Restoration, SONET-SDH, IP, and MPLS*, 1st ed., ser. The Morgan Kaufmann Series in Networking. Morgan Kaufmann, Aug. 2004.

[17] E. W. Dijkstra, "A Note on Two Problems in Connexion with Graphs," *Numerische Mathematik*, vol. 1, no. 1, pp. 269–271, Dec. 1959.

[18] J. W. Suurballe, "Disjoint Paths in a Network," *Networks*, vol. 4, no. 2, pp. 125–145, 1974.

[19] J. W. Suurballe and R. E. Tarjan, "A Quick Method for Finding Shortest Pairs of Disjoint Paths," *Networks*, vol. 14, no. 2, pp. 325–336, 1984.

[20] A. Nucci and K. Papagiannaki, *Design, Measurement and Management of Large-Scale IP Networks: Bridging the Gap between Theory and Practice*, 1st ed. Cambridge University Press, Dec. 2008.

[21] V. Foteinos, K. Tsagkaris, P. Peloso, L. Ciavaglia, and P. Demestichas, "Operator-Friendly Traffic Engineering in IP/MPLS Core Networks," *IEEE Transactions on Network and Service Management*, vol. 11, no. 3, pp. 333–349, Sep. 2014.

[22] M. Yu, M. Thottan, and L. Li, "Latency Equalization as a New Network Service Primitive," *IEEE/ACM Transactions on Networking*, vol. 20, no. 1, pp. 125–138, Feb. 2012.

[23] L. S. Buriol, M. G. C. Resende, and M. Thorup, "Survivable IP network design with OSPF routing," *Networks*, vol. 49, no. 1, pp. 51–64, Jan. 2007.

[24] A. Mereu, D. Cherubini, A. Fanni, and A. Frangioni, "Primary and Backup Paths Optimal Design for Traffic Engineering in Hybrid IGP/MPLS Networks," in *Proceedings of the 7th International Workshop on Design of Reliable Communication Networks (DRCN 2009)*, Washington D.C. (United States), Oct. 2009.

[25] M. Zhang, B. Liu, and B. Zhang, "Multi-Commodity Flow Traffic Engineering with Hybrid MPLS/OSPF Routing," in *Proceedings of the IEEE Global Telecommunications Conference 2009 (GLOBECOM 2009)*, Honolulu, HI (United States), Nov.-Dec. 2009.

[26] O. Gerstel, "The Age of Multi-Layer Networking," in *Proceedings of the Asia Communications and Photonics Conference 2013 (ACP 2013)*, Beijing (China), Nov. 2013.

[27] T. A. Feo and M. G. C. Resende, "Greedy Randomized Adaptive Search Procedures," *Journal of Global Optimization*, vol. 6, no. 2, pp. 109–133, Mar. 1995.

[28] L. Kleinrock, "The Latency/Bandwidth Tradeoff in Gigabit Networks," *IEEE Communications Magazine*, vol. 30, no. 4, pp. 36–40, Apr. 1992.

[29] J. Edmonds and R. M. Karp, "Theoretical Improvements in Algorithmic Efficiency for Network Flow Problems," *Journal of the Association for Computing Machinery*, vol. 19, no. 2, pp. 248–264, Apr. 1972.

[30] P. Pavon-Marino, *Optimization of Computer Networks – Modeling and Algorithms: A Hands-On Approach*, 1st ed. Wiley, May 2016.

[31] "GitHub: Net2Plan Project Source Code," [Last accessed: January 2017]. [Online]. Available: https://github.com/girtel/Net2Plan

[32] R. Ramaswami and K. N. Sivarajan, "Design of Logical Topologies for Wavelength-routed Optical Networks," *IEEE Journal on Selected Areas in Communications*, vol. 14, no. 5, pp. 840–851, Jun. 1996.

[33] P. Pavon-Marino *et al.*, "Offline Impairment Aware RWA Algorithms for Cross-Layer Planning of Optical Networks," *Journal of Lightwave Technology*, vol. 27, no. 12, pp. 1763–1775, Jun. 2009.

[34] S. Orlowski, R. Wessäly, M. Pióro, and A. Tomaszewski, "SNDlib 1.0 – Survivable Network Design Library," *Networks*, vol. 55, no. 3, pp. 276–286, May 2010.

[35] R. S. Cahn, *Wide Area Network Design: Concepts and Tools for Optimization*, 1st ed. Morgan Kaufmann, May 1998.

[36] A. Markopoulou, G. Iannaccone, S. Bhattacharyya, C.-N. Chuah, Y. Ganjali, and C. Diot, "Characterization of Failures in an Operational IP Backbone Network," *IEEE/ACM Transactions on Networking*, vol. 16, no. 4, pp. 749–762, Aug. 2008.

[37] J. M. Simmons, *Optical Network Design and Planning*, 2nd ed., ser. Optical Networks. Springer, 2014.

[38] R. Hülsermann, M. Gunkel, C. Meusburger, and D. A. Schupke, "Cost modeling and evaluation of capital expenditures in optical multilayer networks," *Journal of Optical Networking*, vol. 7, no. 9, pp. 814–833, Sep. 2008.

[39] J. T. Moy, "OSPF Version 2," RFC 2328 (Internet Standard), Internet Engineering Task Force, Apr. 1998.

[40] J.-L. Izquierdo-Zaragoza, J.-J. Pedreno-Manresa, and P. Pavon-Marino, "Maximizing IP Fast Rerouting Coverage in Survivable IP-over-WDM Networks," in *Proceedings of the 41st European Conference on Optical Communication (ECOC 2015)*, Valencia (Spain), Sep.-Oct. 2015.

PLACE PHOTO HERE

Jose-Juan Pedreno-Manresa (S'15) received his B.Sc in Telecommunications Engineering in 2014, from the Universidad Politécnica de Cartagena, where he is currently working toward his Ph.D. degree in the Department of Information and Communication Technologies. His research interests include SDN and NFV techniques for multilayer network orchestration.

He is a member of IEEE and IEEE Communications Society.

PLACE PHOTO HERE

Jose-Luis Izquierdo-Zaragoza (S'12, M'16) received his M.Sc. in Telecommunications Engineering and Master in Information and Communication Technologies in 2010 and 2011, respectively, from the Universidad Politécnica de Cartagena, Spain, where he is currently working toward his Ph.D. degree in the Department of Information and Communication Technologies.

In 2016 he joined Aire Networks, Spain, where he leads the research strategy of the company as the coordinator of the R&D&I Department. He has authored or co-authored more than 20 journal and conference papers. His current research interests include planning and operation of multilayer networks, software-defined networking (SDN), network function virtualization (NFV) and 5G.

He is a member of IEEE, IEEE Communications Society and OSA.

PLACE PHOTO HERE

Pablo Pavon-Marino received an M.Sc. degree in telecommunication engineering in 1999 from the University of Vigo, Spain. In 2000, he joined the Universidad Politécnica de Cartagena, Spain, where he is an associate professor in the Department of Information and Communication Technologies. He received his Ph.D. degree from this university in 2004, and an M.Sc. degree in Mathematics in 2010. His research interests include the planning and optimization of communication networks.