Computer Networks 62 (2014) 152–161

Contents lists available at ScienceDirect

Computer Networks

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

Multi-MetaRing fairness control in a WDM folded-bus architecture

Andrea Bianco [a,*], Davide Cuda [b], Jorge M. Finochietto [c]

[a] Dip. di Elettronica e Telecomunicazioni, Politecnico di Torino, Italy
[b] Cisco systems, Lausanne, Switzerland
[c] Universidad Nacional de Córdoba, CONICET, Argentina

ARTICLE INFO

Article history:
Received 8 March 2013
Received in revised form 1 October 2013
Accepted 3 October 2013
Available online 4 December 2013

Keywords:
WDM
Fairness
Metro architecture

ABSTRACT

The paper deals with fairness issues in a slotted, single-hop, WDM (Wavelength Division Multiplexing) optical architecture, based on a folded bus topology, previously proposed as a broadband access system or as a metro network. The peculiar fairness problem arising in this folded bus based architecture is addressed and an extension of the MetaRing protocol to the WDM scenario, named Multi-MetaRing, is proposed. Feasible Multi-MetaRing strategies are defined and analyzed. Both fair access and high aggregate network throughput can be achieved with a low complexity distributed access protocol by properly handling node access through all WDM channels.

1. Introduction

The bandwidth demand in access systems and metropolitan networks is steadily increasing as a consequence of the Internet usage growth, and of the introduction of new bandwidth-hungry applications. Nowadays, several alternatives, mostly based on packet-switched technologies, to ADSL access and to legacy SONET/SDH in the metro section, are being proposed by vendors.

PONs (Passive Optical Networks) are gaining a significant segment in the broadband access market. They exploit the large bandwidth of optical fibers to reduce complexity of the network infrastructure and of ONUs (Optical Network Units) user interfaces. Packets travel in a single-hop fashion in the optical domain between user interfaces and a central office, with no real optical or electronic switching of packets. However, scalability is constrained by the limited reach and the single-channel operation. The appearance of hybrid TDM–WDM (Time Division Multiplexing–Wavelength Division Multiplexing) PONs increases the complexity of the central office, where all the network traffic is concentrated and processed even for ONU to ONU communications. On the metro segment, RPR (Resilient Packet Ring) and metro Ethernet are two well established and increasingly deployed technologies, both relying on electronic packet switching. However, electronic solutions do not scale well due to the need of electronically processing the whole network bandwidth in each node [1]. As such, there is need to find alternative solutions, based on optical technologies to cope with the increasing traffic requirements.

To cost-effectively sustain larger bandwidths, several optical switching alternatives are today available: OCS (Optical Circuit Switching), OBS (Optical Burst Switching) and OPS (Optical Packet Switching). However, OCS lacks the bandwidth granularity needed to flexibly support highly dynamic traffic in the access/metro segment, true OPS (Optical Packet Switching) is still far from practical feasibility, and OBS (Optical Burst Switching) shows unsatisfactory throughput performance due to large burst loss probabilities [2].

In this paper we consider an alternative solution based on passive WDM infrastructures and tunable transceivers. Similar architectures were studied in the past by several

[*] Corresponding author. Tel.: +39 011 0904098; fax: +39 011 0904099.
E-mail addresses: andrea.bianco@polito.it (A. Bianco), davide.cuda@gmail.com (D. Cuda), jfinochietto@efn.uncor.edu (J.M. Finochietto).

1389-1286/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.bjp.2013.10.013

research groups and are also popular as commercial solutions [3–6]. Nodes are typically equipped with few transceivers, each one operating at the data rate of a single WDM channel. Paths between nodes are created by dynamically sharing on a packet-by-packet basis WDM channels, without requiring nodes to process the full network bandwidth as in electronic networks. Tunability at transceivers is required to exploit the fiber bandwidth by temporally allocating all-optical single-hop bandwidth chunks between nodes in the available channels. These architectures permit to design optical networks that ensure bit rate scalability, because no real packet switching in the optical domain is needed. Indeed, packets travel in the optical domain from the source node to the destination node transparently crossing intermediate nodes. This ensures bit-rate scalability by creating very short-lived end-to-end optical connections. In this context, WDM single-hop optical ring networks are considered as a very promising architecture: The optical medium offers huge bandwidth, the ring topology is able to satisfy fault protection and restoration requirements and complex switching in the optical domain is avoided. thus obtaining a cost-effective balance of optics and electronics.

We focus on a multi-channel (i.e., WDM) single-hop optical network providing any-to-any connectivity to a set of user interfaces (called network nodes), each capable of receiving and transmitting the full bitrate of one WDM channel. The network is based on a physical ring to exploit the restoration properties of the topology, but logically operates as a folded bus, to improve network scalability, as described in [7,8]. Indeed, operating on a folded bus prevents the exploitation of the space reuse feature of ring networks, but it permits to reduce transmission impairments due to noise recirculation typical of all-optical ring networks. Furthermore, no active components like SOA are needed to extract packets from the ring, ensuring lower energy consumption and cost. In summary, the considered network combines features of PONs [8] and of metro networks [7], with good scalability and fault tolerance properties. It may be suited also to broadband optical interconnection systems, e.g., as a switching fabric to interconnect line-cards in a packet switch architecture or to interconnect processors and storage units in a data center, although we do not focus on this scenario in the paper.

Both ring and folded bus topologies introduce unfairness in node access opportunities. This paper introduces a fair access protocol exploiting the previously proposed MetaRing protocol [9]. The design of fairness protocols in a WDM network imposes new challenges. Indeed, since nodes are typically equipped with one transmitter only, a suitable protocol must regulate access not only on a single channel but to different WDM channels to ensure good overall network performance. Furthermore, the techniques adapted to WDM rings [10] as extensions of fairness protocols devised for electronic networks [11] (e.g., MetaRing, ATMring) cannot be directly used in this context. Indeed, in this paper we adapt and extend some of the solutions proposed for WDM rings to the WDM folded bus scenario, because the folded bus topology creates rather different unfairness phenomena as better explained later. As a final observation, we seek for a fully distributed solution, disregarding any centralized control scheme, thus without imposing any constraint on input traffic.

The paper is organized as follows. In Section 2 we describe the considered network architecture and system model. In Section 3 we introduce the fairness problem arising in this network and discuss adaptation of the MetaRing protocol to the specific network architecture. Next, in Section 4 we propose extensions of the MetaRing protocol to the WDM scenario, discussing two different policies that can be used to achieve high throughput and fairness. Simulation results are discussed in Section 5. Finally, we draw conclusions in Section 6. Preliminary results were presented in [12]. The main novel contributions of this paper are: (i) a more refined performance analysis, (ii) the proposal of an analytical model that well approximates some significant protocol parameters, (iii) the delay analysis, and (iv) the study of network scalability in terms of number of nodes and network physical size.

2. Network architecture

We consider a specific WDM optical packet network, physically based on two counter-rotating rings, as sketched in Fig. 1. This architecture, named WONDER, was proposed in [7]. N nodes share W wavelengths. Since typically $N > W$, several nodes receive data from the same wavelength (for instance, in Fig. 1 node i and j receive on λ_x while node l and k receive on λ_y and λ_z, respectively).

Differently from traditional bidirectional ring networks, one ring is used for transmission only (called TX Ring in Fig. 1), while the other one is used for data reception (RX Ring in Fig. 1). To provide connectivity between the two rings, a folding point (loop-back fiber) is needed, where data are moved from the TX Ring to the reception path. Transmitted packets travel from the source node toward the folding point in a first ring traversal, are moved to

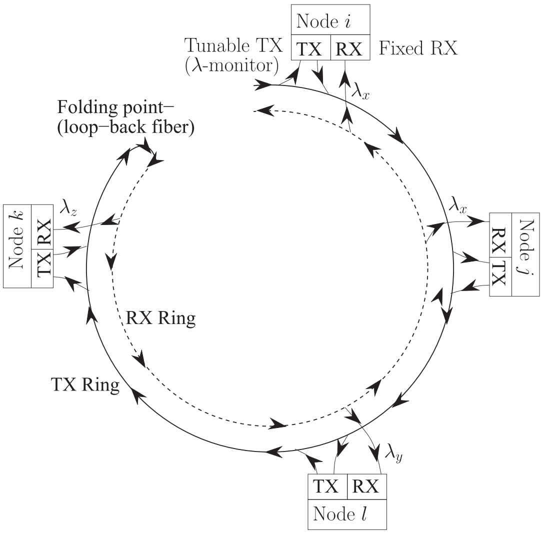


Fig. 1. Considered network architecture: TX Ring and RX Ring are connected through a loop-back fiber, originating a folded bus logical topology.

the reception path, and are received during a second ring traversal. Overall, the physical ring topology with the folding point is equivalent to a folded bus. The folded bus topology prevents the possibility of exploiting the slot reuse feature of ring networks that increases network throughput. However, it avoids the issue of optical signal recirculation typical of ring networks, thus simplifying the node architecture [7].

Observe that the folding point can be created on a dynamically selected node exploiting an Optical Switch (OSW) [13]. Thus, in case of a single fault, a new node can be selected to move the folding point to the proper position in the ring, thus preserving the resilience property of ring topologies.

The network is synchronous on all channels, and time-slotted. The node at the head of the bus generates a synchronization signal on a dedicated wavelength, as discussed in [14], to signal to all other nodes the time slots starting time. Slots propagate on the bus and, at a given time, different slots are available to each access node. The time a packet takes to traverse one ring from the first node to the last node (i.e., half of the folded bus) is measured in time slots, and it is referred to as the network PT (Propagation Time). During a time slot, W slots are available, one for each wavelength channel, for fixed-size packet transmissions.

We assume that nodes are equipped with a single fastly (i.e., on a packet-by-packet basis) tunable transmitter and a fixed burst-mode receiver tuned to one among the W available wavelengths. This permits to keep the node electronic complexity under control although employing more transceivers can provide better performance. Nodes exploit WDM to partition the traffic directed to disjoint subsets of destination nodes: Each subset includes the destinations whose receivers are tuned to the same wavelength. TDM allows to partition traffic between receivers tuned on the same wavelength. Nodes tune their transmitters to the receiver's destination wavelength, and establish a temporary single-hop connection lasting one time slot. Due to the single transmitter architecture, a node can transmit at most one packet per time-slot selecting one among the W available slots. We define as first (last) node, the node at the head (tail) of the bus. A node i is upstream (downstream) to node j when node i is closer to the first (last) node than node j. Thus, the term upstream refers also to the node position with respect to data propagation direction.

Access decisions exploit a channel inspection capability, named λ-monitor, similar to the carrier sense functionality in Ethernet [15]. The channel inspection permits to detect which wavelengths were not used by upstream nodes in each time slot. In similar architectures, an equivalent functionality is obtained by means of a control channel, in which the busy/free slot state information is updated and propagated among nodes. Collisions are avoided by giving priority to the upstream nodes, i.e., to in-transit traffic, Thus, packets are selected for transmission only if the wavelength leading to the destination is free in the current time slot.

In addition, a VOQ (Virtual Output Queued) electronic queue architecture is adopted to avoid the HoL (Head of the Line) blocking problem [16]. Each node keeps separate FIFO queues, each one storing packets for a different destination, or for a different set of destinations (e.g. all the nodes receiving on the same wavelength). In the case of the proposed network, where usually there is more than one node receiving on the same wavelength ($W < N$), there is no performance difference between adopting a queue for each destination (N queues) or a queue for each channel (W queues), because the HoL blocking depends on the channel access opportunity and not on the specific destination node sharing the considered channel. For simplicity and without loss of generality, we adopt one queue per channel.

One issue in these architectures is the receiver-to-channel allocation, which has been discussed in [17]. The optimal allocation policy depends on many factors, including the traffic matrix, if known, and the adopted optimality criterion. As an example, a uniform receiver allocation well matches uniform traffic. If the traffic matrix is not uniform, several "good" allocations may exist. An often adopted heuristic criteria is to seek for an allocation that first balances the load on all channels, and, then balances the transmitter load on each channel, subject to the constraints imposed by the traffic matrix. In this paper, we focus on the fairness properties of the access protocol and assume the receiver-to-channel allocation as given.

3. MetaRing protocol

A problem common to ring and bus topologies is the different access priority given to network nodes depending on their position along the ring/bus. Referring to Fig. 1, it is easy to see that an upstream node can "flood" a given wavelength, reducing (or even blocking) the transmission opportunities of downstream nodes competing for access to that channel, leading to significant fairness problems [18].

Among all available fairness schemes, we selected the MetaRing protocol, due to its good performance in terms of both throughput and delay [11] and to its adaptability to the WDM scenario [10]. The MetaRing protocol was originally proposed to address fairness in ring networks where a single channel is available. In MetaRing, a control signal or message, called SAT (from SATisfied), is circulated cyclically from node to node, normally in the upstream direction with respect to the data flow. SAT transmissions can take place in a dedicated control channel (out-of-band SAT transmission) or using the data channel (in-band SAT transmission). A node forwarding the SAT is granted a transmission quota Q. The node can transmit up to Q packets before the next SAT reception. When a node receives the SAT, it immediately forwards the SAT to the upstream node only if it is satisfied, i.e., if no packets are waiting for transmission, or Q packets were already transmitted since the previous SAT reception. If the node is not satisfied, the SAT is held at the node until the node becomes satisfied. Thus, SATs are delayed only by nodes suffering throughput limitations, and SAT rotation times increase with network load, saturating to $N \times Q$ in heavy overload.

Quota values must be selected with care. To provide full bandwidth to a single node in a ring, the quota Q must be at least equivalent to the time needed by the SAT to return to the node (i.e., equal to the number of data slots contained in the ring). Obviously, also FIFO buffering requirements scale with the quota value. When the quota Q and the buffer size are sufficiently large, MetaRing guarantees 100% throughput: no slots are left unused if the total traffic offered to the network is not smaller than the available bitrate over the transmission medium. Unnecessarily large quotas increase buffering requirements, access delays and traffic burstiness.

If a single-channel folded bus topology instead of a ring is assumed, some further issues arises. First, the value of Q must be larger than in the ring case: Indeed, each time a SAT is forwarded on the folded bus, PT slots are needed on average to reach the next node in the SAT circulation cycle. Therefore, the quota Q must be set equal to at least $N \times PT$, to avoid network under-utilization when only one node is active. Queue lengths must increase accordingly, as previously explained, and fairness enforcement times increase as well. Second, due to the folded bus topology, under overloading uniform traffic, only the last active node delays the SAT. Indeed, the last node has the worse access opportunity because all other nodes are upstream. As a consequence, when the first node forwards the SAT to the last one, all other nodes have already renewed their quotas and typically have also begun transmission. Thus, the last node receives the SAT and delays it until the channel becomes free. In overload conditions, the SAT is delayed until all nodes run out of quota. Since the SAT propagates in the upstream direction, each node releases the SAT and is able to transmit (on average for PT time slots) until it is flooded by the traffic transmitted by the upstream node who has renewed its quota. As a result, when the SAT returns back to the last node, all nodes have some residual quota, depending on the node position on the bus. The average residual quota is equal to Q–PT. Therefore, in overload the SAT will be delayed for $N \times (Q$–$PT)$ slots. Only when the last node exhausts its quota, the SAT is released and forwarded to upstream nodes. However, since all upstream nodes are satisfied (i.e., they ran out of quota), they simply forward the SAT with no delay until the SAT reaches the last node, where it is again delayed until satisfaction is achieved.

Fig. 2 depicts the transmission activities observed from node 4, the last node on the folded bus, with $N = 4$ overloaded nodes under MetaRing control. Arrows labeled SAT$_{i \to j}$ pointing to the time line represent SAT arrivals at node j, while SAT$_{i \to j}$ arrows departing from the time line represent SAT transmissions from node i: Boxes over the time line represent data transmitted by the node whose index is enclosed in the box. The timing diagram starts when node 4 releases the SAT. Assuming upstream SAT rotation, node 4 transmits data for the sum of the propagation delay $P_{4\to3}$ from nodes 4 to 3 and of the propagation delay $P_{3\to4}$ from nodes 3 to 4. Indeed, the SAT must reach node 3 ($P_{4\to3}$) and the first packet transmitted by node 3 must reach node 4 ($P_{3\to4}$). At that time, data transmitted by node 3 are observed at node 4 for $P_{3\to2} + P_{2\to3}$ because node 3 was satisfied: Hence, upon reception of the SAT from node 4, node 3 immediately regenerates its quota and forwards the SAT to the next node (node 2) in the SAT cycle. SAT processing times are neglected in this example. After $P_{3\to2}$, node 2 receives the SAT, and starts transmitting, so that node 3 receives node 2 data and must stop transmitting $P_{3\to2} + P_{2\to3}$ time units after having forwarded the SAT upstream. This same behavior is repeated by all upstream nodes, which, after receiving and immediately forwarding the SAT, have a transmission opportunity equal to $P_{k\to k-1} + P_{k-1\to k}$ for node k. Instead, Node 1 can freely make use of its full quota. After completion of data transmission at node 1, node 2 (and the downstream nodes in turn) complete their quota, i.e., node k transmits for a time equal to $Q - P_{k\to k-1} - P_{k-1\to k}$. Node 4, the last node on the folded bus is the only one holding and delaying the SAT because it is not satisfied. If the nodes are evenly spaced on the folded bus (or on the physical ring), $P_{i\to i+1} = P_{nn}\,\forall i$ and $P_{i\to i-1} = P_{nn}[1 + 2(N - i)] + P_f$, where P_{nn} is the node-to-node propagation delay, and P_f is the delay in the folding from the transmission to the reception bus after the last node; the difference $P_{i\to i-1} - P_{i+1\to i}$ is equal to $2P_{nn}$. The amount of data transmitted by nodes immediately after receiving the SAT depends on the position on the bus, but it is equal on average to PT.

4. Multi-MetaRing protocol

To extend the MetaRing protocol to a multi-channel WDM network, we propose the Multi-MetaRing protocol that makes use of W SATs, each SAT controlling the traffic on a different wavelength channel. In the remainder of the paper we denote by SAT$_w$ the SAT associated with channel w ($w \in \{1,2,\ldots,W\}$). We assume that SATs are transmitted on a dedicated out-of-band control channel, to avoid contentions with data packets. MetaRing extensions to WDM rings were already proposed [10]. However, folded bus topologies impose new challenges due to the fact that,

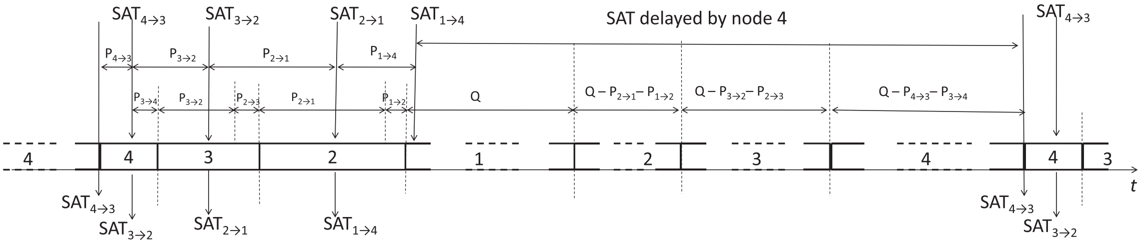


Fig. 2. Timing of SAT circulation on a folded-bus topology.

differently from ring topologies, a node is in the same position when accessing network resources on all the available channels. Furthermore, larger quotas are needed to cope with larger average propagation delays. Thus, previously proposed solutions cannot be directly re-used.

As in the case of a single channel network, the value of Q, for each channel, must be chosen to be larger than $N \times PT$ to avoid throughput limitations when a single node is transmitting on a specific channel. Thus, we assume that nodes are always assigned a quota $Q = N \times PT$. Since W SATs circulate on the network, a node could delay more than one SAT to have equal opportunities to transmit on all channels. However, this would typically deteriorate network throughput, since nodes are equipped with only one transmitter, and if more than one channel is blocked by the same node, slots may be left empty. As a consequence, proper SAT retention policies must be defined; we consider here two possible policies, named Hold-SAT and Release-SAT policies.

4.1. Release-SAT policy

The rationale behind the RSAT (Release-SAT) policy is to hold only one SAT in a node to avoid the throughput loss due to the single transmitter node architecture. Thus, at most one SAT at a time is delayed by a node. Priority is given to the already held SAT: If a node is already holding SAT$_i$ when SAT$_j$ is received, SAT$_j$ is forwarded with no delay. However, to obtain throughput fairness, the node needs to renew its residual quota on channel j by increasing the available quota by Q. This process is referred to as *quota cumulation*. As in the original MetaRing, a node keeps a SAT until it is fully satisfied on the associated channel. Under uniform overloaded traffic conditions, SATs will be alternatively delayed by the last W nodes, who will release SATs only when satisfied.

Although this solution improves network utilization by avoiding the retention of multiple SATs in a single node, it implies that nodes can cumulate a quota larger than Q. As a consequence, fairness is achieved on time windows larger than $N \times Q$ slots (differently from what happens in a single wavelength network). Furthermore, node queues need to be larger, because they must temporarily buffer packets on channels where quota is being cumulated. It is obviously important to check that the quota cumulation process does not diverge.

4.1.1. Modeling the quota cumulation process

To evaluate the quota cumulation behavior we propose a simple discrete-time Markov model. The model considers a uniform overloading traffic scenario on all channels, as the Multi-MetaRing protocol regulates node access only when the network is highly loaded. Indeed, at light loads, nodes receiving SATs are, on average, always satisfied, because their queues are almost always empty due to the low input load and easy channel access. As such, SATs are, on average, not delayed.

The proposed model estimates the amount of quota cumulated by the last node on the bus to properly dimension the node queue size. For simplicity, we focus on the behavior of the last node on the bus, which has the lowest access probability, and on a given channel w. Fig. 3 shows the discrete-time chain associated with the quota cumulation process of the last node on channel w. Each state of the chain represents the number of times a quota Q has been cumulated, i.e., the number of times a quota Q has been added to the current quota value since the last holding of SAT$_w$. This number is also equal to the number of times the last node receives but does not delay SAT$_w$. More in details, when the last node receives SAT$_w$ while holding a SAT associated with another channel, a transition from state i to state $i+1$ occurs. Finally, a transition from state i to state 0 occurs when SAT$_w$ is retained by the last node on the bus, i.e., it is received while the node is not holding any other SAT.

Let P_{SAT} be the probability that the last node is holding a SAT on a single channel network. Under uniform overloaded conditions, the SAT rotation time is equal to $N \times Q$. The last node delays a SAT for a time equal to $N(Q - PT)$, because by the time the SAT is retained, each upstream node has renewed its quota to Q and transmitted on average PT slots. Thus, $P_{SAT} = \frac{Q-PT}{Q}$ and, being $Q = N \times PT, P_{SAT} = \frac{N-1}{N}$.

In a multichannel network, the last W nodes can retain SATs. Thus, the same relationships hold for the last W nodes with some minor modifications. We focus on the worst case which is represented by the last node. We make the simplifying assumption that the probability of delaying a specific SAT on a given channel w is simply equal to $P_{SAT_w} = \frac{P_{SAT}}{W}$. As such, the probability of holding at least a SAT is equal to $P_{SAT_{any}} = 1 - (1 - P_{SAT_w})^W$ and the probability of not holding SAT$_w$ is $P_{SAT_{\overline{w}}} = P_{SAT_{any}} - P_{SAT_w}$.

Let us now evaluate the transition probabilities p and q. The transition probability p represents the probability that the last node starts the quota cumulation process on channel w and it is equal to the probability of delaying a SAT different from SAT$_w$. Thus:

$$p = P_{SAT_{\overline{w}}} \tag{1}$$

Note that p is equal to 0 when $W = 1$, because no quota cumulation process exists in a single channel network. On the other hand, p increases with increasing W: the larger the number of SATs circulating in the network, the higher the probability to start the quota cumulation process.

To compute q, the probability that the last node continues to cumulate quota on channel w, we focus on $1 - q$, the probability of stopping the quota cumulation process, which is the probability that the last node delays SAT$_w$. The last node delays SAT$_w$ only if it is not delaying any

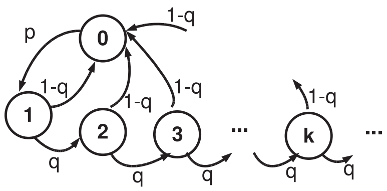


Fig. 3. Markov chain describing the last node quota cumulation process on a given channel w.

SAT and if SAT_w is not delayed by the other last $W-1$ nodes. Thus, q is given by:

$$q = 1 - \left(1 - P_{SAT_{any}}\right) P_{SAT_{\overline{w}}}^{W-1} \tag{2}$$

which increases as W increases. Indeed, as the number of SAT circulating in the network increases, it becomes harder to delay a specific SAT.

Finally, by solving the Markov chain, we obtain the probability π_0 that the node is not cumulating quota on channel w:

$$\pi_0 = \frac{1-q}{1-q+p} \tag{3}$$

and the probability π_k that the node has cumulated k times a quota Q is

$$\pi_k = \pi_0 \times p \times q^{k-1} \ \forall \ k > 0 \tag{4}$$

The average cumulated quota converges to:

$$E\left[\widehat{Q}\right] = \sum_{k=1}^{\infty} Q \times k \times \pi_k = Q \times \frac{p}{(1-q+p)(1-q)} \tag{5}$$

which is finite because $|q| < 1$ and p is finite. This implies that, in our network, the quota cumulation process is bounded to a finite value for a given N. Note that, for increasing N, Q increases as well. Thus, also $E\left[\widehat{Q}\right]$ increases.

4.2. Hold-SAT policy

The HSAT (Hold-SAT) policy states that nodes can hold (delay) more than one SAT at the same time. Thus, if a node receives SAT_j while it is already delaying SAT_i, it holds also SAT_j, if it is not satisfied on channel j. As a result, a node can hold up to W SATs. Similarly to the single-channel scenario, the last node on the folded bus will delay all SATs to access all channels under uniformly overloaded traffic conditions. However, more complex activity patterns can be observed in the multi-channel case with respect to the dynamics shown in Fig. 2 for the single-channel case.

When using the HSAT policy, the Multi-MetaRing protocol ensures a good level of fairness in a relatively short time window, equal to the SAT rotation time, roughly equal to $N \times Q$ slots in overloading uniform traffic. However, this policy can limit throughput performance. Indeed, when the last node delays more than one SAT, being equipped with only one transmitter, it may leave some slots empty. To mitigate this problem, it is important to establish the strategy that nodes must follow to schedule packet transmissions when several SATs are held.

Since the longer the queue occupancy the larger the difficulty in accessing the corresponding channel, nodes can implement a *longest queue* strategy, named in the paper HSAT-LONG, to select packets for transmission among the queues corresponding to channels for which the node is holding the SAT. The longest queue choice is quite common in many systems including optical ring architectures [10] and input-queued switches, and normally provides good performance. If the HSAT-LONG strategy is adopted, a node holding W SATs, with queues approximately of the same length, will delay the SATs until (i) all queues are empty or (ii) the quota on all channels is exhausted. As a result, the maximum delay experienced by all SATs is approximately equal to WQ slots, while the maximum number of slots used for transmission is Q. Thus, the network utilization under uniform overload traffic is temporarily reduced by a factor proportional to $\frac{1}{W}$.

More precisely, when using the HSAT-LONG strategy, in uniform overload, all SATs are blocked by the last node, and released approximately at the same time (i.e., in adjacent time slots). Hence SATs tend to synchronize and visit network nodes at the same time. This introduces a correlation among the activities on the W channels; for example, the first W nodes transmit in parallel and exhaust their quotas on the W available channels almost at the same time. Only when all quotas at the first W nodes are exhausted, the next set of W nodes start transmitting in parallel until quotas are available. Assuming that W is an integer divisor of N, after some time the last W nodes start transmitting in parallel. However, since the residual quota on all channels is increasing from the first to the last node, after some time the first of the last W nodes exhausts its quotas, so that only $W-1$ node transmit in parallel, leaving unused slots. This process is repeated for $W-2, W-3, \ldots$ nodes, until the last node transmits alone to exhaust its quotas. Since the difference between the residual quotas of adjacent nodes is twice the node-to-node propagation delay P_{nn} when node are evenly spaced on the physical ring, it can be shown that the number of unused slots on all channels is equal to $2P_{nn}[1+2+\ldots+(W-1)]$, corresponding to $W(W-1)P_{nn}$ in total, or $(W-1)P_{nn}$ on each channel. Thus, the maximum throughput in uniform overload, when nodes are evenly spaced and W is an integer divisor of N, can be easily shown to be equal to:

$$TH_{max} = \frac{N \times Q}{N \times Q + (W-1) \times P_{nn}} \tag{6}$$

where P_{nn} is the node-to-node propagation delay.

Although some throughput penalties may be acceptable in the optical network context where a large bandwidth is available, throughput losses can be reduced by minimizing node SAT retention time. For this reason, a lowest quota strategy, named HSAT-LOW, is proposed to enhance throughput performance: Among the queues associated with channels for which the node is holding a SAT, the queue associated with the lowest residual quota is selected for transmission. to minimize the SAT retention time. The queues associated with delayed SATs are served sequentially, in almost strict priority ordering, minimizing SAT retention times, thus partly avoiding the synchronization among SATs.

5. Simulation results and analysis

In this section we present performance results obtained by simulation mostly considering a reference network with $W = 4$ wavelengths and a total of $N = 16$ nodes. Node receivers are uniformly distributed among channels. The distance between two adjacent nodes is about 18 km, i.e., 90 µs. Thus, the ring PT is 1.45 ms when $N = 16$. We keep fixed the distance among consecutive nodes. This implies that when the number of nodes N increases, PT also increases. Slots last 1 µs, corresponding to a packet size of about 1250 bytes at 10 Gbit/s. The quota Q is set equal to $Q = N \times PT$. Each node keeps W separate FIFO queues, one for each channel. The size of each node queue is equal to Q when the HSAT policy is adopted, while it is equal to $N \times Q$ when the RSAT policy is selected, according to the findings in Section 5.1, where these values are shown to be able to ensure fairness. As a consequence, queue sizes scale with network size in terms of number of nodes N. Simulation runs exploit a custom simulation environment in C language. Statistical significance of the results is assessed by running experiments with an accuracy of 1% under a confidence interval of 95%.

Two different traffic scenarios are considered: uniform traffic and unbalanced traffic. In the uniform traffic pattern, the whole capacity of the network is equally shared by all nodes. In the unbalanced traffic pattern, nodes are partitioned into two subsets, named server and client. The server subset contains only a single node, named *server*, positioned at the head of the bus to provide a worst-case access scenario. All other nodes belong to the client set. The server transmits at a high rate, equal to the capacity of one wavelength, with equal probability to the other $N - 1$ nodes acting as clients. The remaining network capacity is shared by client nodes; each client transmits $\frac{1}{W-1}$ of its traffic toward the server and the remaining traffic to the other $N - 2$ clients with equal probability. To balance the load on all wavelengths, one wavelength is dedicated to transmit to the server, while clients are equally split among the other $W - 1$ wavelength channels.

The permanent overload traffic case is also considered, not because real networks permanently operate in this regime, but because this scenario may be representative of network behavior under transient overload situations due to a sudden traffic increase or to network reconfiguration.

In most figures, the RSAT policy is plotted using a square, the HSAT-LOW using a circle, the HSAT-LONG using a triangle.

5.1. Uniform traffic scenario

We start considering the normalized throughput (ranging from 0 to 1) achieved by the Multi-MetaRing protocol, as a function of the number of network nodes N, under a uniform traffic scenario when the network is overloaded. Recall that increasing the number of nodes implies increasing the network size, and, as a consequence, node quota and queue size. Fig. 4 shows that the RSAT policy is able to achieve a throughput equal to 1, independently of the network size. On the contrary, the HSAT policy performance depends on the adopted strategy. If the lowest quota scheduling strategy HSAT-LOW is adopted, performance is very close to the one of the RSAT policy, because the protocol is able to desynchronize SAT retentions in the last node. However, if the longest queue scheduling strategy HSAT-LONG is adopted, the achieved throughput is slightly reduced, especially for a small number of network nodes. Indeed, since the longest queue strategy equalizes queue lengths, all the SATs are held and released almost simultaneously by the last node. Hence, the single transmitter bottleneck at the last node introduces a throughput penalty. This penalty is less evident when the number of nodes increases, because the negative effect on throughput due to the last node inability of filling up all the available slots becomes, in percentage, less significant. Throughput reduction is more evident as the number of channels increases (compare white and black marks in Fig. 4), but the absolute performance loss is not dramatic, being smaller than 10%. Overall, the Multi-MetaRing protocol scales well with increasing network size (both in terms of number of nodes and physical distance).

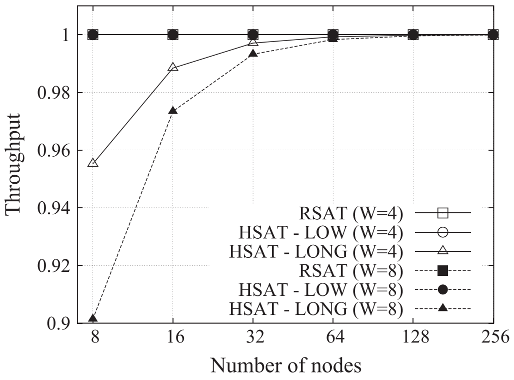
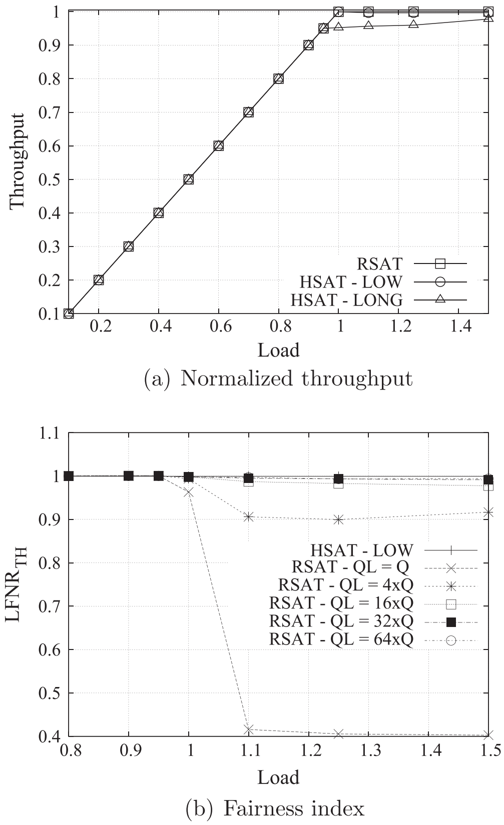


Fig. 4. Comparison of the throughput of the different strategies for different network sizes with offered load = 1 for $W = 4$ and for $W = 8$.

Fig. 5(a) shows the throughput versus the offered load achieved by the different policies when $N = 16, W = 4$. All policies apart from HSAT-LONG achieve a throughput close to 1. Fig. 5(b) shows the fairness index achieved by all the protocols. The HSAT-LONG strategy is not reported because it does not permit to reach the maximum throughput. We plot the ratio between the throughput of the last node and of the first one, labeled as the LFNR (Last-First Node Ratio). A LFNR value close to one represents a fair throughput division, while the lower the value the higher the unfairness. For the RSAT policy, we consider different queue lengths to illustrate the need of longer queues to achieve fairness. The HSAT-LOW strategy is able to ensure fairness in a single cycle, i.e., every $N \times Q$ slots. All nodes transmit Q packets in a cycle on each channel. Thus, a queue length equal to Q is sufficient to achieve fairness. On the contrary, if the RSAT policy is adopted, the level of fairness depends on the node queue length, which must be carefully selected. As described in Section 4.1, a node must be able to buffer an amount of packets equal to the maximum achievable quota. Since the last node cumulates quota until it is able to hold the SAT regulating the traffic on the channel, fairness might not be achieved deterministically in $N \times Q$ slots, but only with a certain probability, equal to the recurrence time $\frac{1}{\pi_0}$ of state 0 in the Markov chain model of Section 4.1.1. Note that a small queue size would imply significant performance losses.

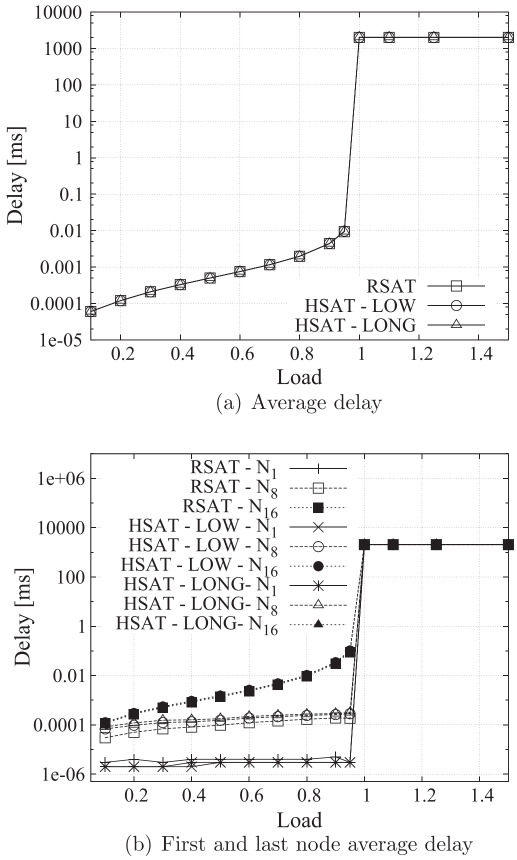
Fig. 6 shows the average delay, including queuing and access delays but not propagation delays, for the different policies. Minor delay differences can be observed. Fig. 6(b) shows the delay for the first, middle and last node on the bus. Multi-MetaRing ensures throughput fairness, but not



(a) Normalized throughput

(b) Fairness index

Fig. 5. Performance of the different MetaRing policies for $N = 16$ and $W = 4$ under uniform traffic.



(a) Average delay

(b) First and last node average delay

Fig. 6. Delays of the RSAT and HSAT strategies under uniform traffic.

delay fairness, because it does not introduce any access control if not in overloaded conditions. Indeed, for a lightly loaded network, all SATs are immediately released because queues are empty most of the time. However, even if the delay difference between the first and last node is very significant, observe that (i) the absolute value of the average delay is fairly small, (ii) the last node delay is close to the average node delay, implying that only the first few nodes on the bus experience a delay reduction, whereas most other nodes face very similar access delays.

5.2. Unbalanced traffic scenario

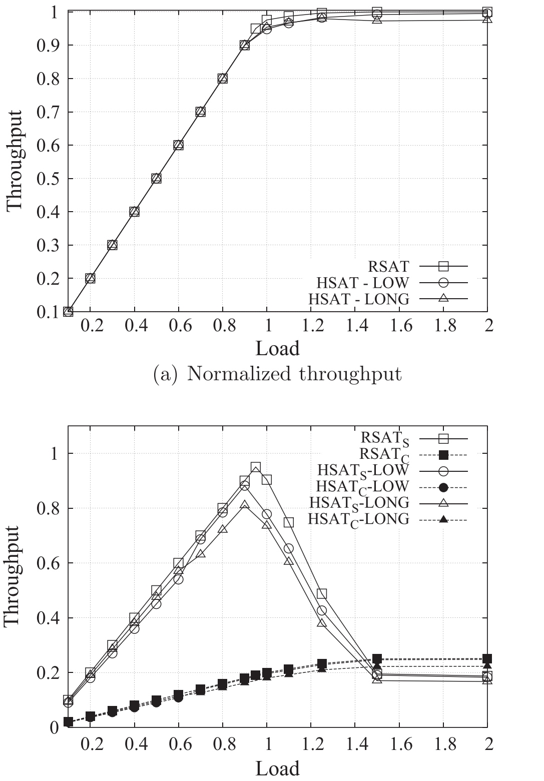
We complete the analysis of the Multi-MetaRing protocol by considering its performance under the unbalanced traffic scenario. Fig. 7(a) shows the average normalized throughput as a function of network load, whereas Fig. 7(b) plots the throughput at the server and at the client node positioned at the end of the bus. All strategies provide high aggregate throughput, the RSAT strategy slightly outperforming the other ones. When the network is in deep overload, the network behaves like under uniform traffic, according to the MAX–MIN fairness paradigm. Fig. 7(b) highlights the MAX–MIN fairness throughput behavior: As the network load increases, the server and the client throughput converge to the same value.

5.3. RSAT cumulation process

Fig. 8(a) shows a comparison between the simulated and the theoretical average quota, varying the number of network nodes N for $W = 4$ under overloading uniform traffic. Recall that we have a constant node-to-node distance. As such, a larger number of nodes implies a larger network size.

As the number of network nodes increases, the average cumulated quota increases, as predicted in Section 4.1.1. Indeed, even though the SAT circulation time increases with the network dimension, also the time SATs are delayed by the W nodes at the end of the bus increases. Thus, it becomes more difficult to stop a SAT on a channel where there is a large quota cumulated. However, for large value of N, this probability converges to a finite value, as proved in the Markovian model. Furthermore, as W increases, the probability of cumulating quota increases accordingly. Since there are more SATs circulating into the network, it becomes harder to stop the right SAT.

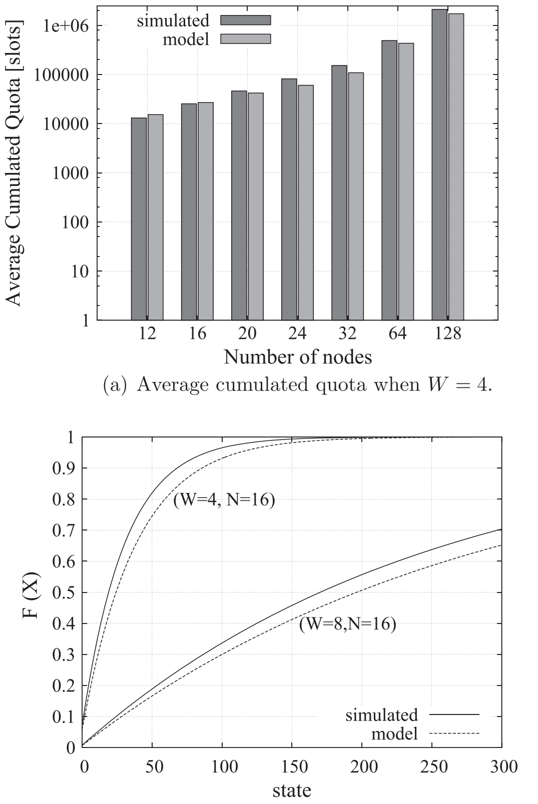
Fairly large queues are needed to sustain the quota cumulation process. Fig. 8(b) shows the simulated and the theoretical cumulative distribution function of the quota cumulation process when $N = 16, W = 4$ and when $N = 16, W = 8$. The proposed model well approximates the results obtained by simulation.



(a) Normalized throughput

(b) Client/Server throughput under unbalanced scenario

Fig. 7. Multi-MetaRing performance under the unbalanced scenario.



(a) Average cumulated quota when $W = 4$.

(b) Quota CDF when $N = 16$ for $W = 4$ and $W = 8$.

Fig. 8. Multi-MetaRing with RSAT policy.

6. Conclusions

We discussed fairness issues arising in a WDM network with *N* nodes and *W* wavelengths, based on a folded bus topology, where nodes are equipped with a single transceiver and *W* electronic queues. We proposed two extensions of the MetaRing protocol to a WDM scenario, named RSAT and HSAT policies, which exploit *W* control signals, named SATs, to ensure throughput fairness on channel access.

The RSAT policy proved to be able to achieve the best performance both under uniform and unbalanced traffic scenarios. However, it presents throughput unfairness if node queues are not large enough. Indeed, nodes must be equipped with a large amount of memory (many times the value of the quota *Q*) to ensure fairness. This creates a scalability issue in the queuing architecture and may also increase the energy consumption. On the contrary, the HSAT-LOW policy is able to ensure fairness in a relative short term cycle ($N \times Q$ time slots), with shorter queues, and achieves performance comparable with the one of RSAT if using a lowest quota scheduling among queues. For these reasons, the HSAT Multi-MetaRing seems the best candidate to control throughput fairness in the WDM network under study.

The minor throughput losses of HSAT-LOW should not be considered as a major issue in the context of optical networks, where bandwidth is easily available and the most critical issues are balancing electronic versus optical complexity, and achieving scalability in network control.

Acknowledgments

We wish to acknowledge Fabio Neri contribution to this paper. Fabio suddenly passed away in April 2011 and we wish to dedicate this paper to his memory.

References

[1] A. Bianco, T. Bonald, D. Cuda, R. Indre, Cost, power consumption and performance evaluation of metro networks, J. Opt. Commun. Network. 5 (1) (2013) 81–91.
[2] P. Pavon-Marino, F. Neri, On the myths of optical burst switching, IEEE Trans. Commun. 59 (9) (2011) 2574–2584.
[3] http://www.intunenetworks.com/home/shape-up/core_innovation/.
[4] http://www.ist-mains.eu/.
[5] D. Chiaroni et al., Demonstration of the Interconnection of Two Optical Packet Rings with a Hybrid Optoelectronic Packet Router, in: ECOC 2010, Torino, Italy, 2010.
[6] B. Uscumlic, I. Cerutti, A.Gravey.P. Gravey, D. Barth, M. Morvan, P. Castoldi, Optimal dimensioning of the WDM unidirectional ECOFRAME optical packet ring, Photonic Netw. Commun. 22 (3) (2011) 254–265.
[7] A. Antonino, A. Bianco, A. Bianciotto, V. De Feo, J.M. Finochietto, R. Gaudino, F. Neri, WONDER: a resilient WDM packet network for metro applications, Opt. Switch. Network. 5 (1) (2008) 19–28.
[8] A. Bianco, D. Cuda, J.M. Finochietto, F. Neri, M. Valcarenghi, WONDER: a PON over a folded bus, in: IEEE GLOBECOM 2008, New Orleans, US, 2008.
[9] I. Cidon, Y. Ofek, MetaRing – a full–duplex ring with fairness and spatial reuse, IEEE Trans. Commun. 41 (1) (1993) 110–120.

[10] M. Ajmone Marsan, A. Bianco, E. Leonardi, F. Neri, S. Toniolo, MetaRing fairness control schemes in all-optical WDM rings, in: IEEE INFOCOM'97, Kobe, Japan, 1997.

[11] S. Breuer, T. Meuser, Enhanced throughput in slotted rings employing spatial slot reuse, in: IEEE INFOCOM'94, Toronto, Canada, 1994.

[12] A. Bianco, D. Cuda, J.M. Finochietto, F. Neri, Multi-MetaRing protocol: fairness in optical packet ring networks, in: IEEE ICC 2007, Glasgow, UK, 2007.

[13] A. Bianco, D. Cuda, J.M. Finochietto, F. Neri, C. Piglione, Multi-fasnet protocol: short-term fairness control in WDM slotted MANs, in: IEEE GLOBECOM 2006, San Francisco, CA, USA, 2006.

[14] S. Bregni, D. Carzaniga, R. Gaudino, A. Pattavina, Slot synchronization of WDM packet-switched slotted rings: the WONDER project, in: IEEE ICC 2006, Istanbul, Turkey, 2006.

[15] A. Carena, V. De Feo, J.M. Finochietto, R. Gaudino, F. Neri, C. Piglione, P. Poggiolini, RingO: an experimental WDM optical packet network for metro applications, IEEE J. Sel. Area. Commun. 22 (8) (2004) 1561–1571.

[16] N. McKeown, A. Mekkittikul, V. Anantharam, J. Walrand, Achieving 100% throughput in an input-queued switch, IEEE Trans. Commun. 47 (8) (1999) 1260–1267.

[17] A. Bianco, J.M. Finochietto, G. Giarratana, F. Neri, C. Piglione, Measurement-based reconfiguration in optical ring metro networks, J. Lightwave Technol. 23 (10) (2005) 3156–3166.

[18] M. Ajmone Marsan, A. Bianco, E. Leonardi, M. Meo, F. Neri, MAC protocols and fairness control in WDM multi-rings with tunable transmitters and fixed receivers, IEEE J. Lightwave Technol. (1996) 1230–1244.

Andrea Bianco is Full Professor and Vice Head at the Dipartimento di Elettronica e Telecomunicazioni of Politecnico di Torino, Italy. He has co-authored over 160 papers published in international journals and presented in leading international conferences in the area of telecommunication networks. He was technical program co-chair of HPSR (High Performance Switching and Routing) 2003 and 2008, DRCN (Design of Reliable Communication Networks) 2005 and ICC 2010 ONS Symposium. He has been TPC member of several conferences, including IEEE INFOCOM, IEEE GLOBECOM, IEEE ICC, HPSR, ONDM and Networking. He is Area Editor for IEEE/OSA Journal of Lightwave Technology and Elsevier Computer Communications Journal. His current research interests are in the fields of protocols and architectures for all-optical networks and switch architectures for high-speed networks. Andrea Bianco is a IEEE Senior Member.

Davide Cuda is Engineer at CISCO systems, Lausanne, Switzerland. He received his Ph.D. in Electrical Engineering from Politecnico di Torino in February 2009, and he graduated from Politecnico di Torino in November 2005. From August 2007 to August 2008 he visited the Networks Research Laboratory of Prof. Biswanath Mukherjee at University of California, Davis, USA. He was Post-Doc at Politecnico di Torino in 2009, at the Italian National Research Council from January 2010 to March 2011 and at Orange Labs, Paris from March 2012 until March 2013. Before moving to CISCO systems in November 2013, he held a Post-Doc position at Telecom ParisTech, Paris. He participated in several Italian and European projects related to optical networks. His main research interests includes asynchronous scheduling algorithms, power and scalability issues of all-optical switching architectures, and design of distributed control algorithms for all-optical switching fabrics.

Jorge M. Finochietto holds a MS and Ph.D. in Electonics Engineer from Universidad Nacional de Mar del Plata, Argentina, and from Politecnico di Torino, Italy, respectively. He is currently Associate Professor at Universidad Nacional de Córdoba, Argentina (UNC), and he is also Adjunct Researcher at the National Research Council (CONICET) of Argentina. From 2005 to 2007, he was a Post-Doc student at the Telecommunication Network Group of Politecnico di Torino. Since 2007 he is member of the Digital Communications Group at UNC. He has been involved in several national and international research projects in the fields of communication networks. He has co-authored over 50 papers published in international journals and presented in leading international conferences. His research interests are in the field of performance evaluation, high-speed networking and switching, and optical and wireless networks.