Resolving the Fairness Issues in Bus-Based Optical Access Networks

Nizar Bouabdallah, Andre-Luc Beylot, Emmanuel Dotaro, *Member, IEEE*, and Guy Pujolle

***Abstract*—Packet-based optical access ring is becoming a promising solution in metropolitan networks. Its performance depends mainly on how optical resource sharing takes place among the different competing access nodes. This network architecture has mostly been explored with regard to synchronous transmission, i.e., slotted wavelength-division multiplexing (WDM) ring. However, in this paper, we focus on the performance of asynchronous transmission-based networks with variable packet sizes. Analytical models are presented in an attempt to provide explicit formulas that express the mean access delay of each node of the bus-based optical access network. We prove that in such a network, fairness problems are likely to arise between upstream and downstream nodes sharing a common data channel. Furthermore, we show that sharing the channel's available bandwidth fairly but arbitrarily between access nodes, as in slotted WDM rings, does not resolve the fairness problem in asynchronous system. In this regard, we exhibit the inherent limitations of the token bucket access rate-based algorithm once applied to asynchronous transmission bus-based networks.**

To alleviate the aforementioned problem, we device a new strategy called traffic control architecture using remote descriptors. The proposed solution is based on a preventive mechanism to grant access to the shared resource. As illustrated in this paper, the proposed solution alleviates the performance degradation and the resource underutilization, while achieving fairness among bus nodes.

***Index Terms*—Access delay evaluation, bus-based optical access network, fairness control, medium access control (MAC) protocol.**

I. Introduction

OVER THE last decade, networks have been witnessing a perpetual growth in data traffic. This shift, driven primarily by the proliferation of internet, has created a rising demand for robust networks, with increasingly high link capacity and node throughput. Due to the new incumbent challenges, the operators are progressively migrating toward optical core networks taking advantage of the tremendous transmission capacity offered by the optical technology. Thanks to the implementation of wavelength-division multiplexing (WDM) in core networks, the relentless need for more capacity may have been satisfied. However, at the boundaries of backbone networks, especially at metropolitan and local area networks, an efficient solution for transporting and switching huge amounts of data still needs to be found.

In next-generation metropolitan networks, Internet traffic will be marked by three major characteristics. First of all, packet-based data traffic of bursty nature will become prevalent. The second characteristic is more related to the fact, that traffic will fluctuate heavily and randomly, which brings us to the third characteristic. Internet traffic will continue to grow during the next few years up to, and eventually beyond, 1 Tbit/s. Consequently, the architecture of next-generation metro networks must evolve to enable tackling these new challenges. In this regard, three major enabling factors are identified as crucial for the evolution process of metropolitan networks' architecture: optics, packet switching, and protocol convergence.

Indeed, packet switching constitutes the necessary ingredient needed for building bandwidth efficient and flexible networks [1]. Asynchronous transmission, which is more suitable for bursty traffic in comparison with slotted WDM rings, must be jointly addressed in the future. Moreover, to minimize the complexity due to layering overhead (overlay schemes such as Internet protocol (IP) over asynchronous transfer mode (ATM) over synchronous digital hierarchy/synchronous optical network (SDH/SONET), the packet-based traffic must be accommodated directly onto optical resources. In this context, the maturing work on generalized multiprotocol label switching (GMPLS) [2] paves the way for the introduction of unique control plane architecture compliant with any underlying switching and transport technology, while also providing traffic engineering features.

In the metropolitan segment, infrastructures are generally organized over a ring topology. This topology type represents a significant investment for carrier operators, and is currently being upgraded to support WDM features. In such networks, packets are electronically buffered at the access nodes' level and are transmitted all optically throughout the ring. Generally speaking, packets' buffering is not performed at intermediate nodes in order to reduce the number of expensive transceivers and to avoid additional processing in the network. To cope with the requirements of next-generation metro networks, a new architecture named dual bus optical ring network (DBORN) has been proposed. The DBORN architecture will be described in this paper, but for more detailed information the reader is invited to refer to [3].

The performance of metro ring networks depends mainly on how optical resource sharing takes place among the different competing access nodes. So, both collision-free transmission and fair sharing of the common bandwidth among ring nodes must be ensured. A medium access control (MAC) protocol is

Manuscript received February 22, 2004; revised September 22, 2004. This work was supported in part by Alcatel Research and Innovation, Marcoussis, France.

N. Bouabdallah and G. Pujolle are with LIP6, University of Paris VI, Paris F-75015, France (e-mail: nizar.bouabdallah@lip6.fr; guy.pujolle@lip6.fr).

A.-L. Beylot is with ENSEEIHT, F-31071 Toulouse Cedex 7, France (e-mail: Andre-Luc.Beylot@enseeiht.fr).

E. Dotaro is with the Alcatel Research and Innovation, Marcoussis F-91460, France (e-mail: Emmanuel.Dotaro@alcatel.fr).

Digital Object Identifier 10.1109/JSAC.2005.851779

needed to avoid collisions on the individual WDM channels shared among competing nodes. The proposed MAC protocol must consider the case of nonslotted WDM rings. But, even though the MAC protocol enables collision-free transmission over the shared medium, yet it does not address the inherent fairness issues, which are pronounced in the case of shared medium networks.

In fact, since several source nodes share a common channel, one node may grab all the available bandwidth, and possibly starve the downstream nodes competing to access the same channel. Protocols at various levels [such as MAC or call admission control (CAC)] must be introduced to ensure good utilization of the transmission resources and to alleviate fairness problems. We present herein analytical models that aim to illustrate the aforementioned issue. The key behavior metric in such networks is the access delay at each node contending for the shared data medium access. In general, fairness control mechanisms limit the transmission of upstream nodes in an attempt to keep enough bandwidth for downstream stations [4], [5]. Although these schemes may be efficient in the case of slotted WDM rings, yet, they do not perform well in the case of asynchronous transmission-based architectures like DBORN. Indeed, we notice that sharing the common bandwidth fairly but arbitrarily among bus nodes, like in slotted WDM ring, does not resolve the fairness problem within asynchronous system. In this regard, we demonstrate the limitations of the token bucket (TB) access rate-based algorithm, used traditionally to regulate the transmission of each bus node.

As a result, a new bandwidth allocation algorithm must be proposed for the asynchronous systems. The main challenge when designing this new protocol is to be able to reconcile two opposing objectives. First, to control the access as deterministically as possible to ensure acceptable packet latency during service delivery. Second, to allow the nodes to take full advantage of the available bandwidth, thus increasing the ring bandwidth usage efficiency. Hence, the expected benefits are both the optimization of resource usage and quality-of-service (QoS) delivery.

We suggest, in this context, a novel control traffic protocol that aims to solve the fairness issue. The proposed solution, the so-called traffic control architecture using remote descriptors (TCARD), presented briefly in [6], is based on a preventive mechanism when granting access to the resource, i.e., free bandwidth is reserved by each node according to the traffic requirements of its downstream partners. We show also how this new protocol deals with the above limitations, through a performance evaluation study. As illustrated hereafter, the proposed solution relieves the performance degradation and the resource subutilization, while achieving fairness among bus nodes.

The remaining parts of this paper are organized as follows. Section II emphasizes the MAC context including a description of the ring network along with its main features. Analytical models for evaluating the access delay performance of each ring node are developed in Section III. Then, Section IV validates the accuracy of the models by comparing the analytical results with that obtained by means of simulations, and it discusses the effects of unfair access to the data channel. The fairness issues and the TB scheme limitations are presented in Section V.

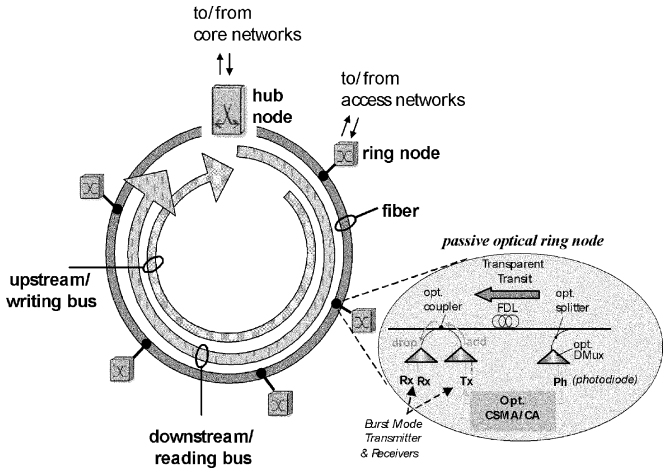

Fig. 1. Overview of DBORN network and node architecture.

Then, Section VI provides a detailed description of the proposed control mechanism TCARD. Section VII demonstrates, through simulation results, how TCARD can achieve fairness and improves the network performances. Finally, conclusions are drawn in Section VIII.

II. Network Architecture and MAC Design

This section describes the DBORN architecture along with its main features and the proposed MAC protocol.

A. DBORN Architecture

DBORN can be described as a unidirectional fiber split into downstream and upstream channels spectrally disjoint (i.e., on different wavelengths). The downstream bus, initiated at the hub node, is a medium shared for reading purposes, while the upstream bus, initiated in the ring nodes, is a multiple access-writing medium. Note that all the transmissions in the ring are performed between the hub and the ring nodes. Therefore, in the downstream direction, DBORN is a point-to-multipoint network, and in the upstream direction it is a multipoint-to-point network.

Let us consider N nodes placed in the unidirectional ring, as shown in Fig. 1. Each node serves one or more access networks. With regard to the direction from the access networks to the feeder ring, the ring node plays the role of a concentrator. Buffered packets are transmitted along the upstream bus toward the hub. In fact, packets travel along the ring without any electro-optic conversion at intermediate nodes. Thus, we deal with the optical transparency propriety of the transit traffic through intermediate ring nodes. The hub terminates upstream wavelengths and electronically processes the packets. According to its destination, a packet is forwarded either into the backbone or through the downstream bus to reach the ring nodes which are being destined. In the latter case, the ring node picks up a copy of the downstream signal, originating from the hub, by virtue of a splitter in order to recover its corresponding packets. Once split, the main signal is no more processed at the node level and it continues its path to serve the other ring nodes, since a wavelength channel could be shared in reception by several nodes. However,

the ring node terminating the copied channel, electronically processes the data packets contained therein, and delivers them to users.

In its straightforward version, each node is equipped with one optical packet add-drop multiplexer (OPADM) and one pair of optical transceivers. The pair of transceivers encompasses a transmitter that is fixed tuned to the node's home transmission wavelength, and a fixed receiver that receives data from the node's home reception wavelength. The OPADM at each node adds the local traffic to the upstream home channel using a combiner, and drops packets contained in the downstream home channel by the means of a splitter. It is important to note that the home wavelength could be shared among several nodes during both reception and emission. This architecture offers several advantages, such as the ability to use a single upstream and a single downstream wavelength, the ability to provision a fractional wavelength capacity to each ring node, a reduced cost of maintenance (due to passive devices use), and both multicast and broadcast capabilities.

B. DBORN Features

As a matter of fact, the spectral separation allows the use of a simple passive structure for the optical part of ring nodes. This architecture inherits the advantages of passive optical networks (PONs) technology viewed by many as an attractive solution to the last mile problem [7]. Passive components (couplers, fiber delay lines) are mature and low-cost components. Here, the novelty relies on the asynchronous optical transmission based on the burst mode transceivers. Such equipments, experimentally validated and already used in PONs, require only slight modifications with respect to synchronous transceivers. Therefore, no limitation is foreseen, either on the physical feasibility or on the complexity of optical components.

Moreover, optical transparency is one of the key features that differentiate DBORN from existing solutions (such as NG-SONET/SDH and RPR). Thus, ring nodes will need transceivers for their local traffic only. In addition, we recall the ability to provide a fraction of the wavelength capacity to each ring node thanks to statistical multiplexing feature, a single wavelength for all upstream nodes, a single head-end receiver at the hub, so on and so forth. Hence, the hub transceiver needs are also reduced.

Indeed, current metro ring networks usually fall into two categories. Either they are based on wavelength circuit switching, keeping SDH/SONET framing to transport different types of client (Ethernet, ATM, IP, SONET, etc.), or they are more packet oriented relying on opaque approaches (switching packets electronically), such as the resilient packet ring (RPR) case. In this regard, switching packets directly in the optical domain, as with DBORN, appears as an attractive solution for future metro networks taking benefits from both packet flexibility and optical transparency. A key advantage of optical packets is to provide a subwavelength granularity to take benefits from statistical multiplexing gain. Moreover, optical transparency reduces the number of transceivers in the network. In this context, DBORN is attempt to develop a simple architecture for high-speed optical access networks, which deals with the aforementioned requirements.

A cost comparison have already been drawn in [3] and [8] between DBORN and most prominent optical ring solutions. Results reveal that DBORN can achieve a significant equipment savings. Based upon the obtained results, we can draw the following conclusions. Compared with the circuit switching approach, the gain is approximately 30%, due to the reduction of the number of transceivers, thanks to statistical multiplexing gain. For instance, SONET rings provide point-to-point circuits among ring nodes. With a dedicated circuit, SONET provides guaranteed bandwidth and delay. However, the primary limitation of SONET rings is bandwidth inefficiency. If all nodes require circuits to all other nodes, an N node ring suffers from the N-squared problem as each node requires $N-1$ circuits. Even for moderate values of N, the total ring capacity is quickly exhausted.

Compared with the RPR solution, the gain can reach 60%, due to the transparency of the DBORN ring nodes. Indeed, unlike a DBORN node, which only deals with its local traffic, each RPR node has to cope with the total traffic on the ring, thus increasing considerably the total cost of transceivers in the network. Recall that RPR networks do not have a true optical node because lightpaths do not pass through nodes, i.e., traffic at each node is processed electronically. In such opaque networks, lightpaths are established only between neighboring nodes. Each node has a single IP router that routes traffic from all the lightpaths. The opaque ring has the advantage of being able to efficiently use the link bandwidth for time-varying traffic thanks to spatial reuse feature. Hence, opaque rings minimize the number of required wavelengths in the network.

However, as nodes do not have optical pass-through, this results in maximum transceiver cost. For instance, in a typical carrier network, each link may have 16 wavelengths, each carrying 10 Gb/s traffic data. Suppose a ring node needs to terminate only one lightpath worth of traffic. In this case, the node would ideally pass through the remaining 15 lightpaths in optical form without processing them. On the other hand, an opaque ring node would require the traffic from all 16 wavelengths to be received, possibly switched through an electronic IP router, and retransmitted. This operation can have a great impact on the network cost. It is worthwhile to note that, in practice, the transceiver cost predominates the cost of the number of wavelengths in the network.

C. MAC Design

In terms of logical performance, the main issue is related to the collision-free packet insertion on a shared writing bus, i.e., the upstream wavelengths. Since the path in transit remains transparent and passive (neither active optical device nor electronic conversion are employed to handle transit frames), no packet is dropped once transmitted on the ring. Hence, traffic control mechanisms are required at the electronic edge of the ring nodes to avoid collision with transit traffic during own data emission.

As stated earlier, we believe that asynchronous transmission allows better use of resources compared with synchronous transmission. Authors in [9] showed the limitations of the time-division multiple-access (TDMA) approach due to the lack of statistical multiplexing. Indeed, a considerable amount

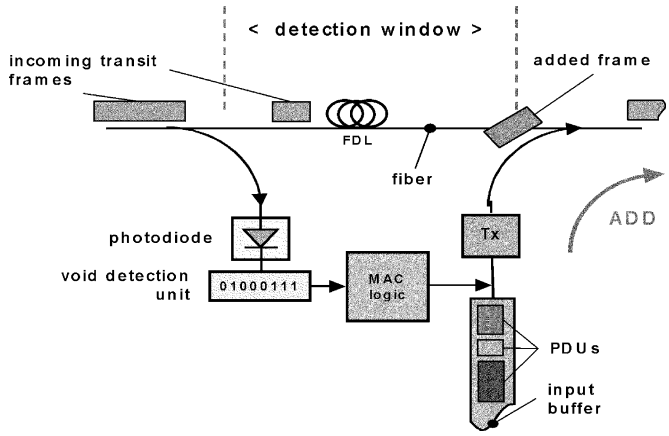

Fig. 2. Schema of the CSMA/CA-based MAC of DBORN.

of bandwidth is wasted due to time slots not being filled to their full capacity. To reduce the cost of metro access networks, it is important to utilize bandwidth effectively. Thus, the future of synchronous transmission, relying on TDMA approach, looks rather grim. Asynchronous transmission, on the other hand, appears to be the preferred choice. Hence, we focus in this paper on contention-based media access protocol (something similar to CSMA) rather than time sharing solution.

We note that ring nodes can not detect a collision once it occurs (because of optical directional proprieties of optical splitter/combiner). The hub however could detect a collision and informs the nodes about collision occurrence by sending a jam signal. However, propagation delays in metro networks, which can exceed 100 Km in length, generally reduce the efficiency of such a scheme. So the well-known access protocol carrier sense multiple access/ collision detection (CSMA/CD) is not appropriate. However, the other extension CSMA/CA (collision avoidance) stands out as the solution of choice. Nevertheless, this mechanism has to be adapted to the optical networks environment.

In a fixed-slotted ring system with fixed-packet size, void (i.e., slot) filling can be carried out immediately upon its detection, since the void duration is a multiple of the fixed-packet size duration. The detected void is, therefore, guaranteed to provide a minimum duration of one fixed-packet length. However in nonslotted ring systems with variable packet length and arbitrary void duration, it is very likely for a collision to occur if a packet is immediately transmitted upon the detection of a void's edge.

To meet these requirements, each ring node must retain the upstream traffic flow within the optical layer, while monitoring the medium activity. So, as shown in Fig. 2, each ring node first uses an optical splitter to separate the incoming signal into two identical parts: the main transit signal and its copy used for control purposes. With regard to the control part, as in [10], low bit rate photodiodes (ph)—typically 155 MHz—are used to monitor the activity of the upstream wavelengths. Once a free state of the medium is detected, the MAC unit measures the size of the progressing void.

To do so, a fiber delay line (FDL) is introduced on the transit path to delay the upstream flow by one maximum frame duration augmented by the MAC processing time. The length of the FDL is slightly larger than the maximum transmission unit (MTU) size allowed within the network, in order to provide the MAC unit with sufficient time to listen and to measure the medium occupancy. The ring node will begin injecting a packet to fill the void only if the null period is large enough (i.e., at least equal to the size of the packet to be inserted). Undelivered data will remain buffered in the electronic memory of the ring node until a sufficient void space is detected. This way, collision free packet insertion on the upstream bus from the add port is ensured.

However, considering only this basic mechanism, head of the line (HoL) blocking and fairness issues arise. A direct result would be performance degradation for ring nodes that are close to the hub node on the upstream bus. Additional flow control mechanisms must be, thus, considered, both at the MAC layer and at the upper layers of edge nodes.

III. Analytical Models

A. Framework

In this section, we will analyze the performance of the network in terms of access delay. The proposed MAC protocol, which is based on CSMA/CA principle, avoids collision between local and transient packets competing to access the shared medium. As described earlier, the MAC protocol detects a gap between two packets on the optical channel, then it tries to insert a local packet into the perceived gap. However, in such an environment, fairness issues could arise.

In the current study, the network is composed of N ring nodes sharing a common medium (e.g., one wavelength) used to contact the hub. Packets arrive to each node according to a Poisson process with the same arrival rate λ in order to exhibit the fairness issue (the analysis presented in this paper can easily be extended to unbalanced traffic conditions). We assume that the transmission time of the packets S forms a sequence of independent identically distributed (i.i.d.) random variables, distributed according to some common distribution function f_S with a mean $E[S]$, a second moment $E\left[S^2\right]$, and a Laplace transform B^*. Moreover, we assume that the length of the packets emitted by the different nodes has the same distribution. The input load ρ_i of a node i ($i = 1, \ldots, N$) is consequently equal to

$$\rho_i = \rho = \lambda E[S]. \tag{1}$$

The aim of this paper is to determine the mean waiting time (or the access delay) of the different nodes $E[W_i]$, defined as the time spent by a packet in the queue i until successfully starting its transmission. Once a packet is emitted, it will not be blocked anymore and will only experience constant delays up to the hub.

We will first study the performance of the first two nodes. An exact model is presented. Approximate analytical methods are then proposed to extend the results to the following nodes, giving upper and lower bounds of the waiting time.

B. Analysis of the First Two Nodes

In order to simplify the analysis, let us primarily consider the first two nodes.

The traffic of the first node has a higher priority to access the medium. The head-of-line packet of the second queue can

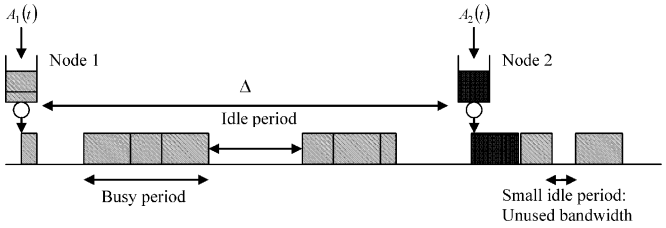

Fig. 3. Activity on the data channel.

only access the channel if the medium is free for a sufficient time period, larger than its transmission time (Fig. 3). So, the emission process of the second node depends on the activity of the first one. The first queue can be simply modeled by an M/G/1 queue. Hence, the waiting time of the first node is given by

$$E[W_1] = \frac{\lambda E\left[S^2\right]}{2(1-\rho)}. \tag{2}$$

So in the remainder study, we will focus on the second queue analysis. This method will be iterated to determine the performance of the other nodes.

In this paper, the "link state" refers to the state of the link when a packet, from a ring node, attempts to access the data channel. The wavelength channel can be in one of two states: free (idle) or occupied (busy). It is obvious that for packets from the upstream node, i.e., node 1, the channel is always idle. However, when packets from the downstream node, i.e., node 2, try to access the channel, the latter can be either free or occupied by upstream traffic. It is important to note in this regard that the state of the medium, as seen by node 2, alternates continuously between an idle and a busy period.

Let $\{A_i(t), t \in \mathbf{R}\}$ denote the arrival process of packets to the queue i.

As the delays due to the propagation from node 1 to node 2 (Δ_1) and to the FDL (Δ_2) are constant, the whole system can be analyzed as a priority queue with preemptive repeat identical (PRI) discipline [11]. The arrival process of packets is defined as, $A(t) = A_1(t-\Delta) + A_2(t)$, where traffic 1 has the higher priority and $\Delta = \Delta_1 + \Delta_2$. The workload for the queue consists of two classes of jobs. The objective is to determine the average waiting time for jobs of each class in the queue. Note that the waiting time of the higher priority class $E[W_1]$ is simply the waiting time in an ordinary M/G/1 queue, as described in (2). Next, we will focus on the waiting time of the class i customers where $i \geq 2$.

Under a preemptive repeat policy, service is interrupted whenever an arriving customer has higher priority than the one in service. The new arrived customer begins service at once. A preempted job will restart service from the beginning as soon as there are no higher priority jobs remaining in the queue. In other words, the preemptive repeat strategy stipulates that the work already done on an interrupted job is lost. In this case, the transmission time of the interrupted packet may be resampled according to the service time distribution after every preemption (preemptive repeat different discipline) or it may be the same as in the first service attempt (preemptive repeat identical discipline). In this study, we adopt the PRI discipline since it coincides with the real behavior of the network.

We can consequently apply the results presented in [11] based on [12]. Let C_i denote the completion time of a class i customer (i.e., the time between starting and finishing service, including the preemption time). Let S_2 be the transmission time of the packet of class 2 that is chosen first. Suppose that $\widetilde{n}$ preemptions occur because of the arrival of $\widetilde{n}$ packets of class 1. Let $I(n)$ be the service time futilely expended due to the nth preemption, and $B(n)$ be the duration of the nth preemption. Note that $I(n)$ is the nth unusable idle period encountered by the packet, while trying to access the data channel and $B(n)$ is the nth busy period of packets of class 1. The completion time C_2 for a packet of class 2 can be written as

$$C_2 = S_2 + \sum_{n=1}^{\widetilde{n}} I(n) + \sum_{n=1}^{\widetilde{n}} B(n). \tag{3}$$

Which leads to

$$E[C_2] = \frac{\overline{X}_{1,1}}{1-\rho}, \text{ with } \overline{X}_{1,1} = \frac{B^*(-\lambda)-1}{\lambda}. \tag{4}$$

The mean waiting time may be derived as follows. In [11], the mean waiting time $E[Z_2]$ is the time spent by a class 2 packet from its arrival until service begins. It does not include the completion time. The mean response time $E[R_2]$ is consequently equal to

$$E[R_2] = E[Z_2] + E[C_2]. \tag{5}$$

As explained before, we refer in this paper to the waiting time, as the time spent by a packet in the queue until its transmission successfully begins. The mean waiting time $E[W_2]$ can be written as

$$E[W_2] = E[R_2] - E[S]. \tag{6}$$

In the case of p traffic classes, we have (7) shown at the bottom of the page. With

$$\rho_p^+ = \sum_{k=1}^{p} \lambda \overline{X}_{k,1} \tag{8}$$

$$\overline{X}_{k,1} = \frac{B^*(-(k-1)\lambda)-1}{(k-1)\lambda}, \quad \overline{X}_{1,1} = E[S] \tag{9}$$

$$\overline{X}_{k,2} = \frac{2\{B^*(-2(k-1)\lambda) - B^*(-(k-1)\lambda) + (k-1)\lambda B'^*(-(k-1)\lambda)\}}{(k-1)\lambda} \tag{10}$$

$$\overline{X}_{1,2} = E[S^2]. \tag{11}$$

$$E[Z_p] = \frac{\sum_{k=1}^{p} \lambda \left(\overline{X}_{k,2}\left(1-\rho_{k-1}^+\right) + \frac{2(B^*(-2(k-1)\lambda) - 2B^*(-(k-1)\lambda)+1)\rho_{k-1}^+}{((k-1)\lambda)^2} \right)}{2\left(1-\rho_{p-1}^+\right)\left(1-\rho_p^+\right)} \tag{7}$$

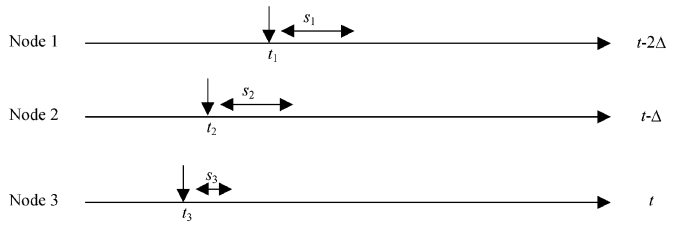

Fig. 4. Example of scenario of packets arrival in three nodes bus.

Solving (7) for $p = 2$, we get (12), as shown at the bottom of the page. We can determine the mean waiting time of the second node queue, which is given by

$$E[W_2] = E[Z_2] + E[C_2] - E[S]. \tag{13}$$

C. Extension to N Nodes

1) An Upper Bound for the Mean Waiting Time: Unfortunately, the previous method cannot be applied to the following nodes. Indeed, in the single priority queue with PRI discipline, the emission time already elapsed on an interrupted job is lost and cannot be used anymore by lower priority jobs $(i+1, \ldots, N)$. However, in reality, if the idle period is not long enough to support the queue i head-of-line packet, the medium remains free and this idle period can be used by downstream nodes.

Let us consider the example depicted in the Fig. 4. All the queues are supposed to be empty at arrival epochs. Let t_i denote the arrival instants of packets Pi on the different queues and s_i their transmission time. In the model with one server under PRI discipline, the server starts serving the arriving packet of the third node (P3). The arrival of a packet in the second node (P2) interrupts the service of P3 still present in the server. This latter returns to the head of its priority class, whereas P2 starts its service immediately. While serving P2, a new packet arrives to the first queue (P1). The service of P2 is preempted and the service of P1 starts at once. Finally, packets of class 2 and 3 have to remain in their queues until the passage of the class 1 packet, in order to try once more to access the shared medium.

However, this is not what happens in reality. Let us now see the real network behavior under the same scenario (Fig. 5). Due to the early arrival of the packet P1 in the first queue, as depicted in Fig. 5(b), the second node detects an idle period insufficient to transmit its packet P2 since

$$t_1 + \Delta - t_2 < s_2. \tag{14}$$

The medium remains free and the left idle period can be used by the downstream node, i.e., node 3. The packet P2 has to wait the passage of P1 and attempts once more to access the medium. Moreover, as

$$t_1 + 2\Delta - t_3 > s_3 \tag{15}$$

the third node detects, thanks to the MAC protocol, a suitable idle period [Fig. 5(c)]. Thus, unlike the analytical model analysis of the system, P3 is sent before the passage of P1 [Fig. 5(d)]. Therefore, the real waiting time of the class 3 customers is lower than the results claimed by the approximate analytical model.

Using this method, it can be shown that the analysis of the system with a single priority queue will lead to an upper bound of the mean waiting time for the downstream node k, where $k > 2$

$$E[W_k] \leq E[W_k^+] = E[Z_k] + E[C_k] - E[S]. \tag{16}$$

Where $E[Z_k]$ is derived using (7) and

$$E[C_k] = \frac{\overline{X}_{k-1,1}}{1 - \rho_{k-1}}. \tag{17}$$

2) Lower Bound for the Mean Waiting Time: Conversely, the following method leads to a lower bound for the waiting time. In each node, the upstream traffic has a higher priority than the local traffic. So, the emission process of the local queue depends only on the activity of the upstream nodes and the profile of busy and idle periods generated by upstream flows. The method consists on aggregating all the upstream traffics in a single flow. The packets of the aggregated flow arrive according to a Poisson process. Then, we analyze each node as a single queue with two traffic classes under PRI priority discipline, where the local traffic has the lower priority.

This approximate analysis leads to an underestimation of the mean response time because it may cause longer busy period duration and consequently longer idle period duration as well. In reality, the free bandwidth seen by a downstream node is much more fragmented than the one generated by the aggregated upstream flow. One then obtains the following results by applying the method of Section III-B to each node k with two flow classes (i.e., upstream and local traffic) with respective arrival rates

$$\lambda_{k-1}^- = (k-1)\lambda, \quad \lambda_k = \lambda. \tag{18}$$

It corresponds to "equivalent loads":

$$\rho_{k-1}^- = \lambda_{k-1}^- E[S], \quad \rho_k^- = \lambda_{k-1}^- E[S] + \frac{B^*\left(-\lambda_{k-1}^-\right) - 1}{(k-1)}. \tag{19}$$

Based on (13) the lower bound of the waiting time is given by

$$E[W_k^-] = E[Z_k^-] + E[C_k^-] - E[S]. \tag{20}$$

$$E[Z_2] = \frac{\lambda\left(E\left[S^2\right] + \frac{2(1-\rho)\left\{B^*(-2\lambda) - B^*(-\lambda) + \lambda B'^*(-\lambda)\right\}}{\lambda^2} + \frac{2\rho(B^*(-2\lambda) - 2B^*(-\lambda) + 1)}{\lambda^2}\right)}{2(1-\rho)(2 - B^*(-\lambda))} \tag{12}$$

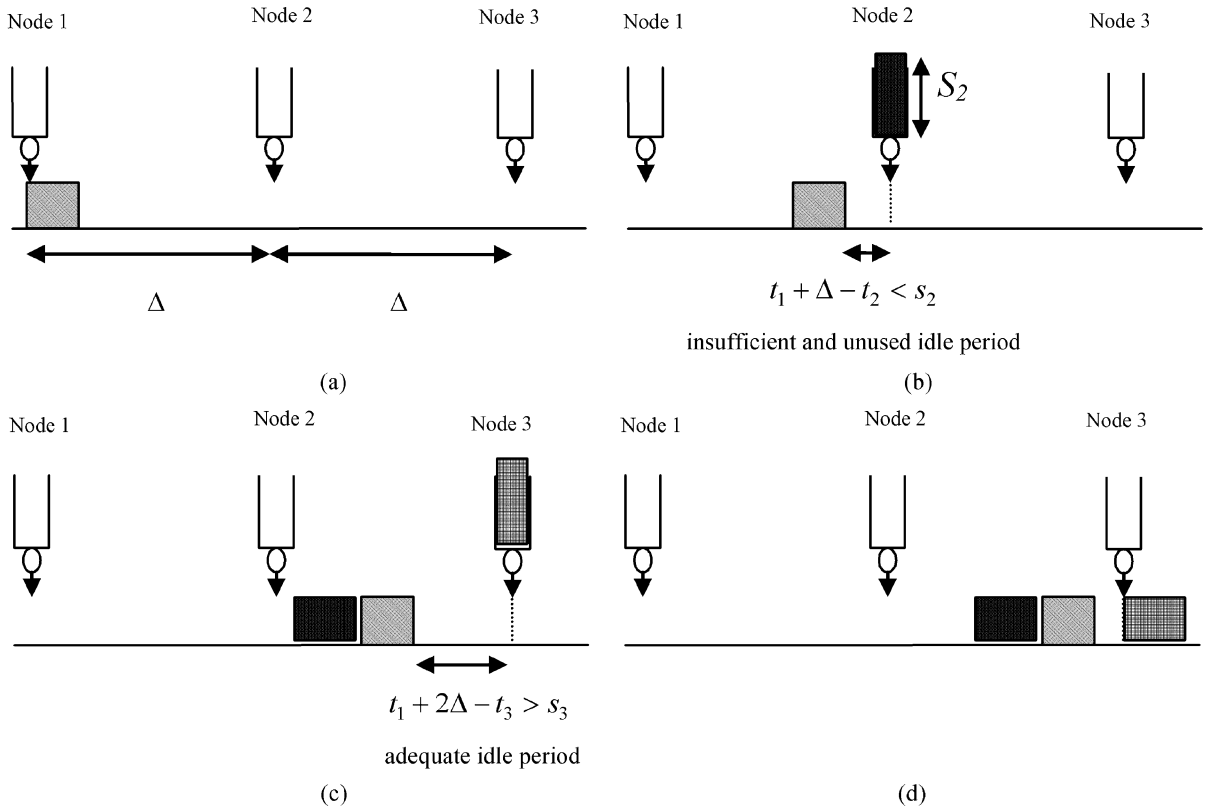

Fig. 5. Evolution of the network state in the time: (a) at $t = t_1 + s_1$; (b) at $t = t_2$; (c) at $t = t_3$; and (d) at $t = t_3 + s_3$.

Respectively, from (4) and (12), we have (21)–(22), as shown at the bottom of the page.

D. Example

Different packet length distributions can be considered. In this paper, we consider packets of variable length (50, 500, and 1500 bytes) more or less representative of the peaks in packet size distribution in Ethernet. Let p_i be the probability of the different packet sizes and d_i the corresponding emission time.

The mean waiting time of the first queue (2), of the second queue (13), and the bounds on the waiting times for the following nodes (16) and (20) can be derived using the following parameters:

$$E[S^k] = \sum_i p_i d_i^k, \quad B^*(s) = \sum_i p_i e^{-sd_i}. \tag{23}$$

IV. Numerical Results and Fairness Issues

To evaluate the accuracy of the proposed analytical models, we compare their results with those obtained from simulations carried out using network simulator 2. In order to meet this purpose, additional modules have been developed to model the DBORN and MAC architectures. In the following, only a subset and a synthesis of the results are presented. In all our simulations, unless otherwise specified, we assume that: 1) all the ring nodes share a common upstream wavelength that runs at 1 Gb/s; 2) the packets arrive according to a Poisson process; 3) the arrival rate of the packets at each node is the same in order to highlight the fairness issues; and 4) all the ring nodes transmit only to the hub. In the remaining part of this paper, the traffic load ρ denotes the average traffic load sourced from every node to the hub. Furthermore, we consider packets of variable length (50, 500, and 1500 bytes) more or less representative of the peaks in packet size distribution in Ethernet. The total traffic volume comprises 50% of 1500 Bytes, 40% of 500 Bytes, and 10% of 50 Bytes packets size.

The analysis results of access delay for the first two nodes are presented in Fig. 6, revealing a perfect match with the simulation results: analytical results practically coincide with the simulation results. We observe the following.

- Under light traffic load, the access delay of the downstream node is bigger than that of the upstream node. As

$$E\left[C_k^-\right] = \frac{B^*\left(-\lambda_{k-1}^-\right) - 1}{\lambda_{k-1}^-\left(1 - \rho_{k-1}^-\right)} \tag{21}$$

$$E\left[Z_k^-\right] = \frac{\left(\frac{2\left(1-\rho_{k-1}^-\right)\left\{B^*\left(-2\lambda_{k-1}^-\right) - B^*\left(-\lambda_{k-1}^-\right) + \lambda B'^*\left(-\lambda_{k-1}^-\right)\right\}}{\lambda^2} + \frac{2\rho_{k-1}^-\left(B^*\left(-2\lambda_{k-1}^-\right) - 2B^*\left(-\lambda_{k-1}^-\right) + 1\right)}{\lambda^2} + \lambda_{k-1}^- E\left[S^2\right]\right)}{2\left(1 - \rho_{k-1}^-\right)\left(1 - \rho_k^-\right)} \tag{22}$$

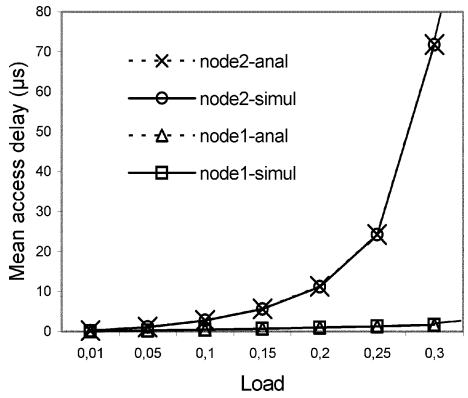

Fig. 6. Mean access delay of the first two nodes with variable-packet size traffic.

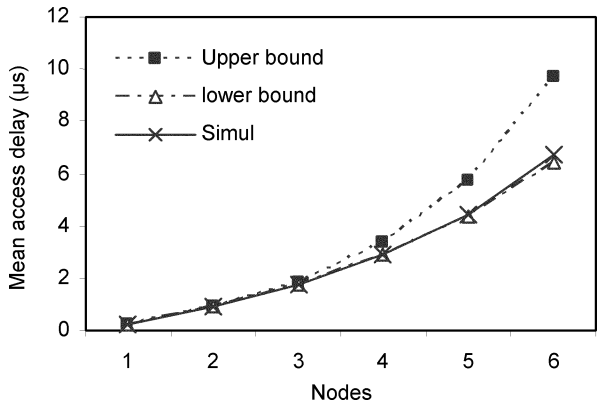

Fig. 7. Mean access delay of the six-node bus with variable-packet size traffic.

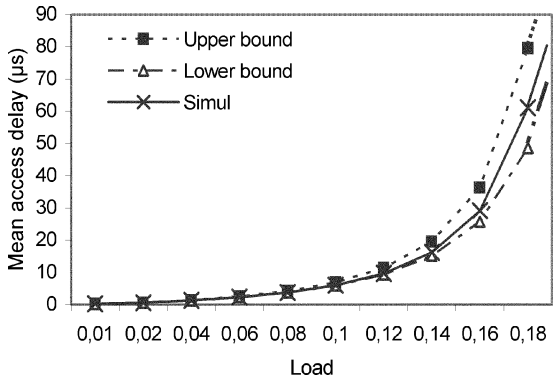

Fig. 8. Mean access delay of the third node of the shared bus with variable-packet size traffic.

a result, the fairness issue is pronounced even under light traffic load.

- Under high traffic load, the performance difference between the upstream and downstream nodes sharing the optical channel increases. The main reason for this fact is that under high load condition the upstream node will grab more bandwidth, thus leaving less capacity to the downstream node.

The analysis results in this special scenario are significant. We observe that even when the upstream node uses a small part of the available bandwidth, the downstream nodes' performance is strongly affected. The fairness issue is always present in shared medium networks. This is mainly due to the lack of organization during the emission process in the network. In fact, the mismatch, between the idle period distribution resulting from the upstream nodes' utilization of the medium and the packets' size distribution of the downstream node, often leads to bandwidth waste, as well as fairness problems with regard to resource access.

Indeed, the poor performance is not due to the saturation of the medium: For instance, the fairness issue is pronounced even when the channel occupancy is no more than 20%. Unfortunately, the available bandwidth is divided mostly among useless idle periods (voids), which can not be filled by new packets emanating from the downstream node. A finer analysis of the problem shows that it results from the combination of two factors.

1) Wasted voids (i.e., voids smaller than the minimum packet size).
2) Inadequacy of voids with respect to the first packet at the head of the first-in–first-out (FIFO) insertion buffer of the ring node.

Once a packet of maximum size is at the head of the insertion buffer, it blocks the emission process until finding an adequate void: this is the well-known HoL blocking problem. Monitoring the distribution of voids on the medium reveals a low probability of finding regular and sufficiently large available room. Thus, sharing the bandwidth fairly but arbitrarily between nodes is not sufficient to ensure satisfactory results. The sharing process must, thus, be done smartly in order to preserve a maximum of useful available bandwidth for downstream nodes. In general, fairness control mechanisms limit the transmission of upstream nodes to keep enough bandwidth for downstream stations. These schemes may be efficient in the case of slotted WDM rings. However, as it appears, they do not perform well in the case of asynchronous transmission-based architectures like DBORN. To bore out our predictions, we evaluate in the next section the performance of the network when the TB rate-based access algorithm is implemented at each ring node.

The analysis results for the general case of six-node bus, depicted in the Fig. 7, emphasize the abovementioned results. The traffic load ρ sourced by each node is 0.05. The access delay of each node is found to increase monotonically when progressing toward the hub. Indeed, the closest nodes to the hub encounter relatively large delays. In contrast, upstream nodes (numbers 1–4), which grab gradually the available bandwidth, are satisfied and benefit from short access delay. We insist that the performance degradation of downstream nodes is not due to the medium saturation since the medium occupation is not beyond 30%. This example exhibits the fairness issue already seen in the particular case of two nodes. The upper and lower bound curves are very close to the simulation result curve. So, the approximate analytical models can achieve high accuracy. But, we make the observation that the bounds become less accurate for the closest nodes to the hub, especially, the upper one. This negligible difference, well below 5% for the lower bound for instance, is expected since the approximations concern more and more nodes when coming near the hub.

The accuracy of the lower and upper bounds is studied further in the Fig. 8, which presents the evolution of the access delay of

node 3 in function of the load ρ. This plot shows that approximate methods are highly accurate when the load of each node is relatively low. However, these models become less accurate with the increase of the load.

In this paper, the exact analytical model handling the case of two-node ring is studied revealing a good match with the simulation results. However, it can be easily extended to more nodes without major modifications but still approximations have to be accounted for. The approximate models are shown to be accurate but we point out that they loss slightly this accuracy with the increase of the load and the number of ring nodes.

V. Token Bucket (TB) Limitations

Usually, a TB is used to police a traffic flow. Such a policer checks and ensures that the traffic flow fits inside a traffic envelope. For a source of traffic the specification of the envelope deals either with the peak rate or the sustainable rate with an associated burst size. These specifications have to be provided when making a reservation with either the resource reservation protocol (RSVP) in the case of IntServ or may be specified in a service level specification (SLS) when using DiffServ. For instance, in the DiffServ model, the customer contracts an SLS with the provider in which the profile of the aggregate traffic is specified.

Although variations exist, there are two basic parameters associated with the operation of a TB: the bucket size b that is the maximum acceptable size of packets' burst; and the token generation rate r, which is the maximum sustainable bit rate. A packet passing through a TB needs to obtain tokens equal to its packet size. Tokens are accumulated at a rate r bits/s up to a maximum equal to the bucket size b. The available tokens are decremented when a packet passes through the TB. If the number of tokens is not sufficient to transmit an arriving packet, then the connection is classified as violating its negotiated throughput. The TB may discard the packet at the entrance of the subnetwork when the algorithm is used for policing purposes. Alternatively, if the algorithm is used for shaping purposes, the violating packet may be buffered at the TB and allowed to enter the subnetwork only when enough tokens are available. In this paper, the TB is used as a shaper.

The simulation results of eight-node bus when TB is enabled are depicted in the Fig. 9. The traffic load ρ sourced by each node is 0.1, i.e., the average rate traffic sent by each node is 0.1 Gb/s. The sustainable bit rate allowed by the TB is set successively to 0.09, 0.1, and 0.11 Gb/s at each node. Besides, the maximum burst size b is set to 10 000 bytes and the capacity of the electronic buffer at each node is set to 1 Mbytes. Results emphasize the above presented expectations and show the inefficiency of the TB scheme. The access delay of each node is found to increase monotonously when progressing toward the hub for both cases: when TB algorithm is enabled or disabled. Indeed, the closest nodes to the hub encounter relatively large delays, which is incompatible with performances needed in metropolitan networks. Meanwhile, upstream nodes (numbers 1 to 4), which grab gradually the available bandwidth, are found to be satisfied and benefit from short access delay.

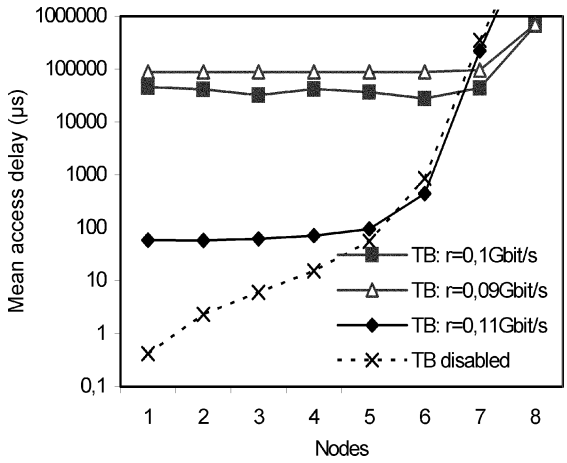

Fig. 9. Mean access delay of the eight-node bus with variable-packet size traffic.

We insist on the fact that the performance degradation of downstream nodes is not due to the medium saturation since the medium occupation does not exceed 80%. This example exhibits the fairness issue already seen and highlights the limitations of TB mechanism to alleviate this problem. Even if the traffic sourced by upstream nodes does not violate the negotiated throughput, it causes unacceptable packet delay to downstream nodes sharing the same channel. Certainly, thanks to the TB algorithm, the free bandwidth (stated in bit/s) allocated to each node is theoretically sufficient to handle its traffic. However, the main issue pertains to the inappropriate distribution of the free bandwidth. This limitation concerns not only the TB algorithm but all the methods based on the arbitrary limitation of upstream nodes emission in order to satisfy downstream ones sharing the same channel.

Hence, we suggest preserving bandwidth (represented by idle periods) by upstream nodes in order to satisfy downstream nodes' requirements in an organized way. A basic rule consists of avoiding random division of the resource, which would lead to inadequacy between idle periods length and the packet size. Therefore, the control mechanism must prevent greedy upstream stations from taking more than their fair share by forcing them to keep idle periods of sufficient size. In other words, the control mechanism required in this context must regulate the distribution of idle and busy periods.

VI. Proposed Protocol: TCARD

To regulate the transmission of a node, we can either use open loop schemes such as the TB algorithm or closed loop schemes based on end-to-end exchange of control messages and back pressure mechanisms. The latter solution involves a large processing overhead per individual packet and requires large buffers at the input node. Furthermore, in a high-speed environment like optical networks, where end-to-end propagation delays are much larger than the transmission time of a typical packet, the effectiveness of such closed loop schemes which relies mainly on some sort of feedback information from the network becomes questionable. In this regard, recently, there has been much emphasis on open loop control schemes, where the

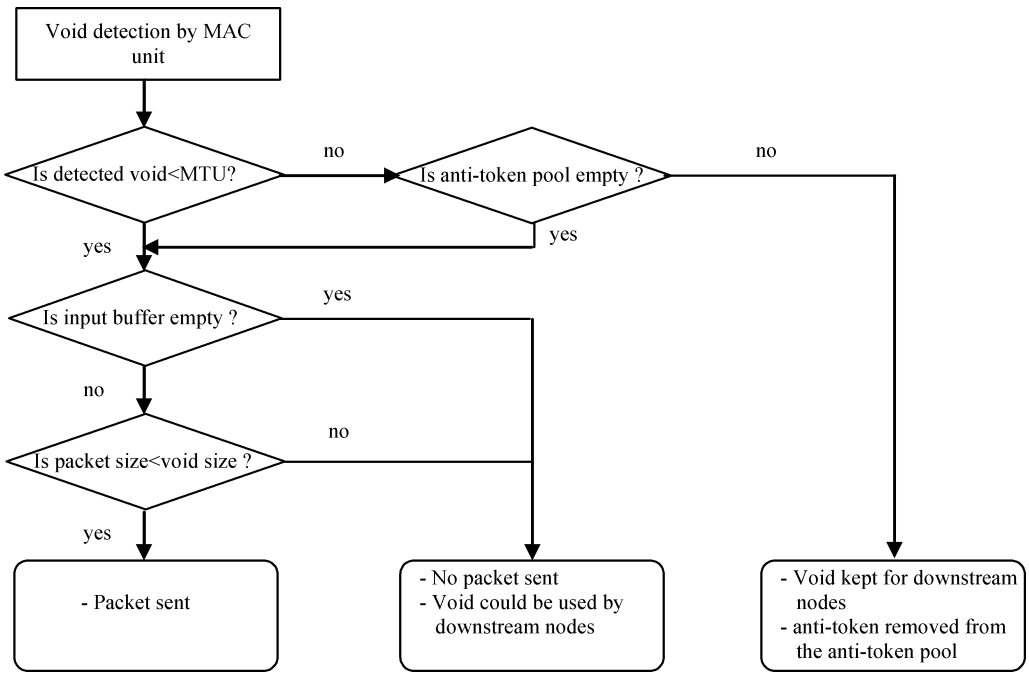

Fig. 10. Decision graph in a TCARD-enabled MAC.

goal in this context is shifted more toward the prevention of unfair optical resources' sharing, rather than toward the attempt of reacting to it. An example of the open loop scheme is the TB rate-based control scheme put into action in the case of slotted WDM ring. However, as explained earlier, this same mechanism presents several limitations when dealing with the fairness problem in the case of asynchronous transmission. Indeed, it is difficult to throttle a source emission and to control at the same time the corresponding idle period distribution when using TB rate-based access control.

To cope with the factors mentioned in the previous paragraph, we suggest a new preventive control mechanism that forces a node to conform to its allocated bandwidth, and it also prevents the random division of the resource capacity. Let us consider N nodes sharing a common unidirectional channel traveling to the hub and let us consider more specifically the ith node of the bus. Unlike the TB algorithm which is based on the ith node traffic descriptor, the TCARD algorithm relies on the specification of the aggregate traffic sourced at the downstream nodes $(i+1,\ldots,N)$. The TCARD mechanism is based on the distribution and the gathering of remote information (descriptors) relative to the traffic requirements contracted by downstream ring nodes sharing the same resource. For each ring node i $(i = 1,\ldots,N)$, the aggregate descriptor reflects the traffic needs of its downstream neighbors. This information is then used to constrain the access of the node i to the medium.

The basic operation of such a scheme is simple. To describe it, let us consider an anti-token pool at the node i, where anti-tokens are generated at fixed time intervals that correspond to the specified average rate of the aggregate downstream traffic. Unlike the TB scheme where each token represents the permission to transmit one bit, each TCARD anti-token prevents the node i from transmission on a detected idle period for a fixed amount of time d. This idle period of size d is reserved for downstream nodes. Following this reservation, an anti-token is removed from the anti-token pool. Packets arriving to the node enter first to the input buffer. Thanks to the MAC protocol, the node listens and measures the void duration present in the medium. If the detected null period is smaller than d, i.e., the anti-token cannot be placed, the void could be used by the node for its own usage purpose as long as the cumulative size of frames to be sent is smaller than the current void size. If the detected null period is equal or greater than d, it is used mainly for the release of one anti-token if the anti-token pool is not empty, otherwise, the void can be used by the node for transmission. The arrival of an anti-token to the pool during the transmission of a packet does not preempt the emission process. The anti-token has to wait the packet service completion in order to be served. The TCARD algorithm is depicted in Fig. 10.

Hence, the main idea behind TCARD lies in the preservation of bandwidth (represented by voids) by upstream nodes in order to satisfy downstream nodes' requirements. However, organized and sensible reservation schemes must be applied to fully benefit from the protocol efficiency. A basic rule consists in avoiding random division of the resource leading to inadequacy between the idle period length and the packet size distributions. Indeed, in order to guarantee for a downstream node the ability to transmit packets of maximum size, the length d of a reserved void (which consumes one anti-token) must be at least equal to the MTU size. On the other hand, the maximum length of an idle period that can be reserved is limited by the FDL length adopted in the MAC design. As stipulated earlier, the FDL length used is slightly larger than the MTU size allowed on the network. Therefore, to fulfill the aforementioned requirements the duration d that has to be adopted must be equal to the time required to transmit the MTU on the shared medium.

To illustrate the TCARD mechanism, we present a simple three-node bus example. The nodes share a common channel

that runs at 1 Gb/s. We assume that the sustainable bit rate negotiated by each node and stipulated in its own SLS is 0.3 Gb/s. We consider traffic of variable packet size where the MTU is equal 1500 bytes. Considering the TCARD scheme, the first node must reserve 0.6 Gb/s in average of available bandwidth for the downstream nodes, i.e., nodes 2 and 3. As explained before, the reserved bandwidth is representative of idle periods of 1500 bytes. Thus, the anti-tokens at node 1 are generated periodically at a rate equal to $(0.6 \cdot 10^9)/(1500 \cdot 8)$ anti-tokens/s. Note that a reserved void can be exploited by a downstream node either to transmit a packet of maximum size or to emit a burst of smaller frames. In addition, similarly to the first node, the second node reserves 0.3 Gb/s of available bandwidth for the third node. The reserved bandwidth is also representative of voids of 1500 bytes.

The proposed method presents two obvious advantages. First, the reservation scheme adopted by each node avoids the random division of the available bandwidth seen by its downstream neighbors. Second, the average rate of traffic transmitted by each node is no more limited by the sustainable rate specified in the SLS as in the TB case. Considering again the previous example, if the first node, for instance, uses only a small part of its available bandwidth, the second node can benefit of its reserved part, as well as of the unused part of the first node. Hence, this approach would allow the excess bandwidth to be used by other ring nodes. Conversely, the TB algorithm limits the transmission of each node to its sustainable rate independently from the current resource usage state. This feature is certainly a major differentiator of TCARD with respect to the TB mechanism. It allows certainly the improvement of the resource utilization.

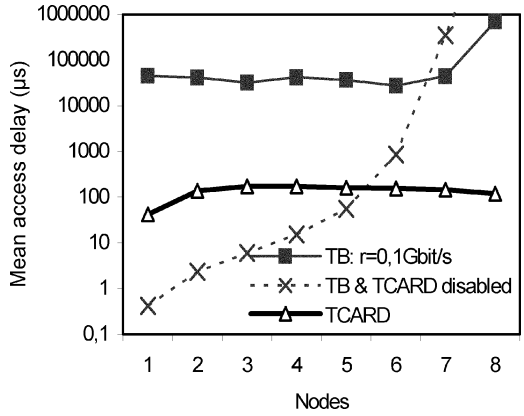

Fig. 11. Mean access delay of the eight-node bus with an input load representing 80% of the medium capacity.

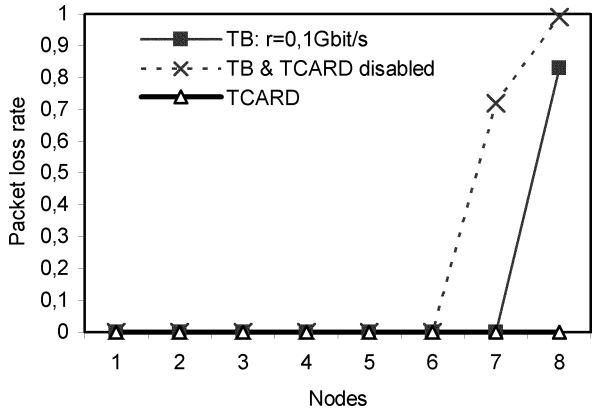

Fig. 12. Packet loss rate of the eight-node bus with an input load representing 80% of the medium capacity.

VII. Performance Evaluation

In this section, we evaluate the performance of the proposed fairness control protocol. We keep the same assumptions used within the simulations presented in Sections III and V. We consider a ring of eight nodes sharing the same wavelength traveling to the hub. Each node receives traffic from the access networks to be forwarded toward the hub at a mean rate of 0.1 Gb/s. Thus, the traffic sourced by all the ring nodes represents 80% of the wavelength capacity. In addition, in this section, we assess the access delay and the packet loss rate (PLR) at each ring node. As explained before, the anti-token generation at each node within TCARD is configured to reflect the average amount of traffic expected at downstream nodes. Moreover, the bandwidth reserved for downstream nodes' use purpose is representative of voids of 1500 bytes in order to comply with packets of maximum size. For instance, the TCARD anti-tokens are generated at the first node at a rate of $(0, 1 \cdot 7 \cdot 10^9)/(1500 \cdot 8)$ anti-tokens/s.

We first compare the performance of each ring node when TCARD is enabled, when the TB algorithm is enabled, and when both TCARD and TB algorithms are disabled.

Fig. 11 depicts the average access delay experienced by packets arriving to each node. Results confirm the limitations of the TB algorithm and highlight the fairness issue already discussed. We point out that the performance degradation, when TCARD is disabled, is not due to the medium saturation since the channel occupancy is below 70%. Recall that the input load is 80%. This difference is simply due to the packet loss resulting from buffer overflow of downstream nodes.

On the other hand, TCARD enables fairness and better use of the resource by sharing efficiently the bandwidth between competing nodes. The mean access delay is around 160 μs for all the nodes. Unlike the TB case, the performance of the downstream nodes when going closer to the hub is not affected by the upstream nodes. In addition, simulations show that TCARD improves the resource utilization, which increases from 70% to 80%. We point out however that the delay recorded at upstream nodes is slightly increased, with respect to the case where no fairness control mechanism is applied, but still remains below 170 μs. This is because TCARD algorithm imposes more constraints on upstream nodes in order to preserve usable bandwidth for downstream ones.

Fig. 12 depicts the PLR at each ring node. As expected, when TCARD is disabled, packet loss occurs more and more when approaching the hub due to the node buffer overflow. Indeed, downstream nodes do not find suitable idle period to transmit their packets. In particular, the loss rate registered at node 8 is above 99% in the absence of any control mechanism and exceeds 80% when TB is enabled. In contrast, with TCARD, no packet loss is recorded in the network. In this case TCARD never incurs loss due to its efficient share of bandwidth among nodes.

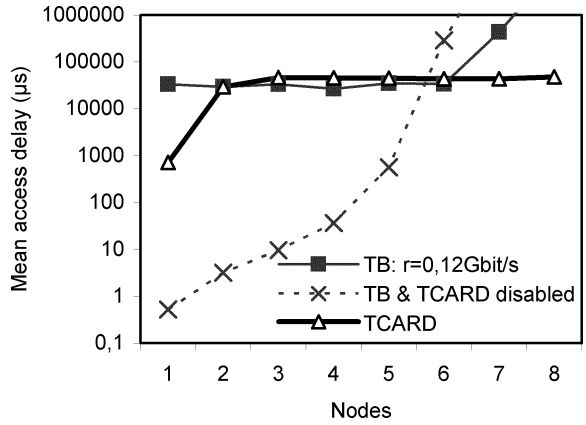

Fig. 13. Mean access delay of the eight-node bus with an input load representing 96% of the medium capacity.

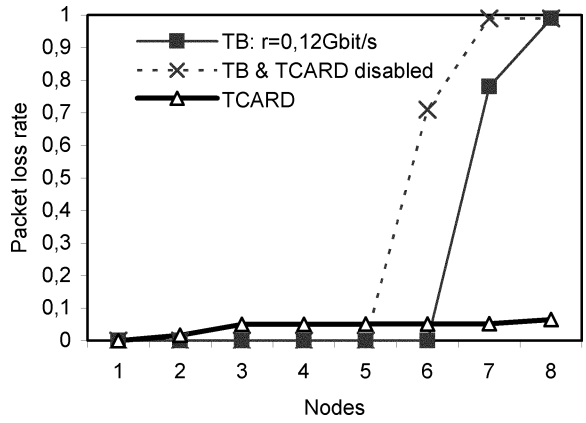

Fig. 14. Packet loss rate of the eight-node bus with an input load representing 96% of the medium capacity.

Figs. 13 and 14 depict, respectively, the access delay and the PLR at each node in an extremely loaded environment. Each node receives traffic from the access networks to be forwarded toward the hub at a mean rate of 0.12 Gb/s. Thus, the traffic sourced by all the ring nodes represents 96% of the wavelength capacity. In this case, TCARD and TB enable nearly the same access delay for upstream nodes (Fig. 13). More precisely, the access delay enabled by TCARD is slightly larger than the one presented by TB. This result is expected, since in these conditions, both TCARD and TB limit each node transmission to its contracted sustainable bit rate. Indeed, there is no excess available bandwidth that can be exploited by TCARD-enabled nodes. Recall that unlike the TB algorithm, which limits the transmission of each node to its sustainable rate independently from the current resource usage state, TCARD allows ring nodes to benefit from the possible excess bandwidth on the shared medium. However, in this particular case, there is no available excess bandwidth since the medium capacity is entirely utilized. In addition, TCARD imposes more constraints to each node in order to preserve useful bandwidth to its downstream neighbors, which explains the minor increase of the access delay recorded at upstream nodes with respect to the TB case. However, when coming closer to the hub, the access delay explodes with TB due to the arbitrary bandwidth reservation. In contrast, with TCARD, the access delay remains unchanged when approaching the hub.

On the other hand, all the ring nodes exhibit a similar loss rate, well below 6%, when TCARD is enabled (Fig. 14). The packet loss is distributed fairly among the ring nodes. Conversely, no packet loss is recorded at upstream nodes when TB is considered. In this case, packet loss occurs more and more when approaching the hub due to the node buffer overflow. Indeed, downstream nodes do not find suitable idle period to transmit their packets. In particular, the loss rate registered at node 8 is above 99%. We point out however that the packet loss rate recorded at upstream nodes is slightly increased, when TCARD is applied, but still remains below 6%. This is because TCARD algorithm imposes more constraints on upstream nodes in order to preserve usable bandwidth for downstream ones.

Moreover, it is meaningful to compare the input buffer utilization at each node among different strategies. Fig. 15 reports the input buffer utilization for the cases where TCARD is enabled, where the TB algorithm is enabled, and where both TCARD and TB algorithms are disabled. The input buffer utilization is plotted as a function of the physical node location and the medium load. Based upon this results, we draw the following conclusions.

- If TCARD is disabled, the input buffer utilization increases with the medium load and when coming closer to the hub. As expected, unfairness problems are always present even when TB algorithm is applied. We notice also that for moderate and heavy load, downstream nodes buffer utilization exceeds 90%. This results in buffer overflow and induces high packet loss rate.
- If TCARD is enabled, the input buffer utilization only depends on the medium load evolution. In other words, it does not depend on the physical node location. Moreover, the input buffer utilization is always below 6% even under heavy loaded medium. We underline, in this context, that upstream nodes input buffer utilization is slightly increased, with respect to the TB case, since TCARD imposes more constraints to upstream nodes in order to preserve useful bandwidth for downstream stations.

Next, we investigate the choice of the void length d that is set intuitively to the maximum packet size. We keep the same simulation scenario. Each node tries to forward 0.1 Gb/s worth of traffic toward the hub. Here, we consider two variants of TCARD: one in which the duration d of a preserved void is set to the transmission time of a packet of 1500 bytes, and one in which d is set to the transmission time of 1000 bytes. In other words, the amount of available bandwidth reserved by each node for its downstream partners is still the same but its distribution changes. For instance, the anti-tokens are generated at the first node at a rate $(0.1 \cdot 7 \cdot 10^9)/(1500 \cdot 8)$ anti-tokens/s when d corresponds to 1500 bytes, whereas the anti-token generation rate is set to $(0.1 \cdot 7 \cdot 10^9)/(1000 \cdot 8)$ in the second case. Recall that d is limited by the FDL size which is 1500 bytes. Simulation results presented in Figs. 16 and 17 show that TCARD has a lower success when each generated anti-token corresponds to a void size of 1000 bytes. The fairness issue is not entirely resolved since the downstream nodes do not find adequate void for

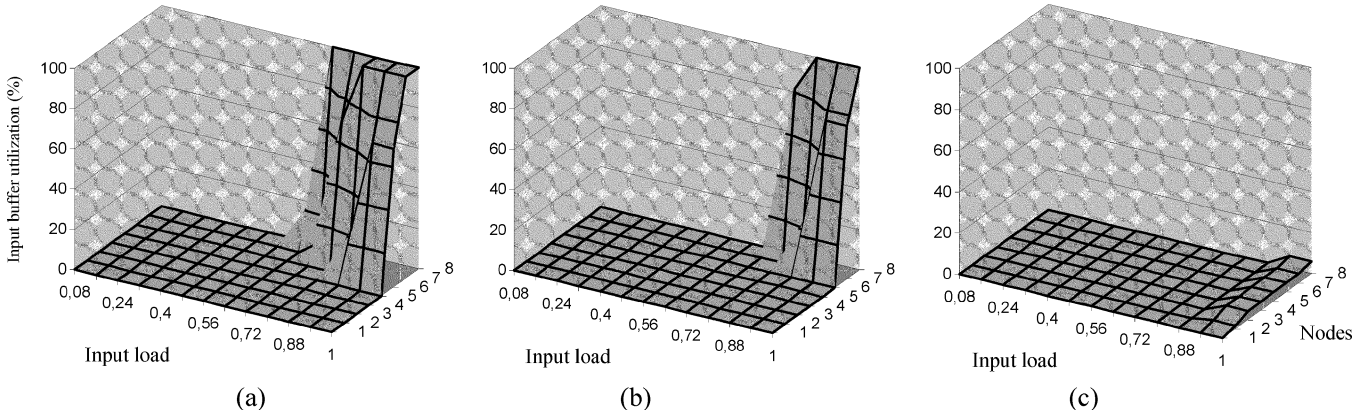

Fig. 15. Input buffer utilization of each ring node. (a) Both TCARD an TB are disabled. (b) TB is enabled. (c) TCARD is enabled.

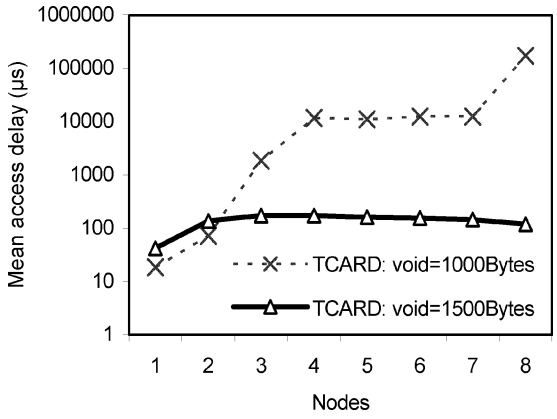

Fig. 16. Mean access delay of the eight-node bus with the two variants of TCARD.

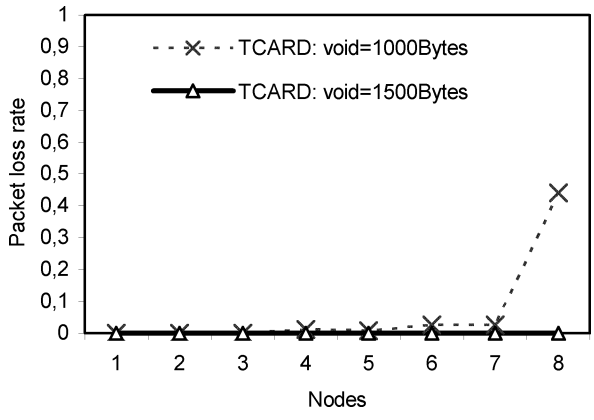

Fig. 17. Packet loss rate of the eight-node bus with the two variants of TCARD.

packets of maximum size. Some nodes (in particular node 8) still suffer from starvation problem. Thus, arbitrate sharing of the bandwidth between nodes is not sufficient to ensure satisfying results. The sharing of the bandwidth must be done smartly in order to preserve a maximum of useful bandwidth for downstream nodes. Consequently, TCARD configuration is a key criterion to achieve good performances.

VIII. Conclusion

This paper provides the analysis of shared bus network's behavior with asynchronous transmission. We analyzed the system performance in terms of access delay required by each node to inject a packet on the shared medium. The analysis results showed that fairness issues are likely to arise between upstream and downstream nodes even under light loads. We observed that sharing the available bandwidth fairly and arbitrarily between nodes does not resolve the fairness problem. In this regard, we exhibited the limitations of the TB rate-based access algorithm. Consequently, an additional flow control mechanism has to be considered, not only to limit the transmission of the upstream nodes but also to organize their emission process. We suggested a new protocol called TCARD, based on a preventive reservation scheme, to alleviate the fairness problem. The reservation process consists in forcing the upstream nodes to ensure enough available bandwidth for downstream nodes' transmission purpose. The simulation results showed how the proposed solution alleviates the performance degradation and the resource subutilization, while achieving fairness among bus nodes.

There are some practical issues that remain to be studied. For instance, in this paper, we have not addressed the ability of TCARD mechanism to deal with traffic of different class-of-service (CoS). As we can see from the TCARD description, the algorithm is based on anti-token generation process. The generation of anti-tokens is representative of idle period profiles expected by downstream nodes. The role of the anti-token generation may differ according to the CoS required by the downstream nodes: typically, the "best effort" anti-tokens generation may reflects the sustainable rate of best effort traffic expected by downstream nodes, while "real-time services" anti-tokens generation can correspond to the peak rate in order to guarantee the service. Therefore, the virtual reservation process could be QoS-aware and can take into account the various service requirements.

References

[1] M. J. O'Mahony, D. Simeonidou, D. K. Hunter, and A. Tzanakak, "The application of optical packet switching in future communication networks," *IEEE Commun. Mag.*, vol. 39, no. 3, pp. 128–135, Mar. 2001.

[2] E. Mannie *et al.*. (2003) Generalized Multiprotocol Label Switching Architecture. [Online]. Available: draft-ietf-ccamp-gmpls-architecture-07.txt

[3] N. Le Sauze *et al.*, "A novel, low cost optical packet metropolitan ring architecture," in *Proc. ECOC*, vol. 4, Amsterdam, The Netherlands, Oct. 2001, pp. 66–67.
[4] M. A. Marsan, A. Bianco, E. Leonardi, F. Neri, and S. Toniolo, "Metaring fairness control schemes in all-optical WDM rings," in *Proc. INFOCOM*, vol. 2, Kobe, Japan, Apr. 1997, pp. 752–760.
[5] J. S. Yih, C. S. Li, D. D. Kundlur, and M. S. Yang, "Network access fairness control for concurrent traffic in gigabit LANs," in *Proc. INFOCOM*, vol. 2, San Francisco, CA, Mar. 1993, pp. 497–504.
[6] N. Bouabdallah, L. Ciavaglia, E. Dotaro, and N. Le Sauze, "Matching fairness and performance by preventive traffic control in optical multiple access networks," in *Proc. Opticomm*, Dallas, TX, Oct. 2003, pp. 424–429.
[7] G. Kramer and G. Pesavento, "Ethernet passive optical network (EPON): Building a next-generation optical access network," *IEEE Commun. Mag.*, pp. 66–73, Feb. 2002.
[8] N. Le Sauze *et al.*, "Optical packet switched metro networks," presented at the ECOC, Copenhagen, Denmark, Sep. 2002.
[9] G. Kramer, B. Mukherjee, and G. Pesavento, "Ethernet PON (ePON): Design and analysis of an optical access network," *Phot. Net. Commun.*, vol. 3, no. 3, pp. 307–319, Jul. 2001.
[10] R. Gaudino *et al.*, "RINGO: A WDM ring optical packet network demonstrator," in *Proc. ECOC*, vol. 4, Amsterdam, The Netherlands, Sep. 2001, pp. 620–621.
[11] H. Takagi, *Queueing Analysis*. Amsterdam, The Netherlands: N. Holland, 1991, pt. I, vol. I, Vacation and Priority Systems.
[12] N. K. Jaiswal, *Priority Queues*. New York: Academic, 1968.

Nizar Bouabdallah received the B.S. degree in telecommunications engineering from Ecole Superieur des Communications (Sup'Com), Tunis, Tunisia, in 2001, and the M.S. and Ph.D. degrees in network and computer science from the University of Paris VI, Paris, France, in 2002 and 2004, respectively.

He joined Alcatel Research Laboratories, Marcoussis, France, in 2002, while working on his Ph.D. degree. He is currently a Visitor Researcher at North Carolina State University. His research interests include optical networking, performance evaluation, network planning and modeling, as well as control and management architectures.

André-Luc Beylot received the Engineer degree from the Institut d'Informatique d'Entreprise, Evry, France, in 1989 and the Ph.D. degree in computer science from the University of Paris VI, Paris, France, in 1993. In January 2000, he received the Habilitation à Diriger des Recherches from the University of Versailles, Versailles, France.

From 1993 to 1995, he worked as a Research Engineer at the Institut National des Télécommunications, Evry, and from 1995 to 1996 at C.N.E.T. (France Telecom Research and Development), Rennes. From September 1996 to August 2000, he was an Assistant Professor at PRiSM Laboratory, University of Versailles. Since September 2000, he has been a Professor at the Telecommunication and Network Department, National Polytechnics Institute of Toulouse (INPT/ENSEEIHT) and is a member of the IRT Team of the IRIT Laboratory. His research interests are performance evaluation of communication networks, especially with regard to high-speed and mobile networks.

Emmanuel Dotaro (M'98) received the M.S. degree from the University of Versailles, Versailles, France, in 1996. He is working towards the Ph.D. degree in computer science from Institut National des Télécommunications, Evry, France, while teaching at the University of Versailles.

He joined Alcatel Research Laboratories, Marcoussis, France, in 1999, where he is now coordinating activities on packet/transport integration in metropolitan networks. His current research interests include performance evaluation and network dimensioning, data, control (GMPLS) and management architectures, and solutions enabling integrated approaches.

Guy Pujolle received the Ph.D. and Thèse d'Etat degrees in computer science from the University of Paris IX and XI, Paris, France, in 1975 and 1978, respectively.

He is currently a Professor at the University of Paris VI. He is Chairman of IFIP Working Group 6.2 on Network and Internetwork Architectures. He is a pioneer in high-speed networking, having led the development of the first gigabit network tested in 1980. He was also a European expert involved in the development of European high-speed networks. He is a cofounder and CSO of QoSMOS and Ucopia Communications. He is an Editor for the *International Journal of Network Management*, *ACM WINET*, and *Ad Hoc Networks Journal*.

Dr. Pujolle is an Editor for the *IEEE Tutorials and Surveys*. He is a Governor of the International Council for Computer Communications.