Traffic Performance of Buffering Strategies in All-Optical Nodes for IP Networks

Stefano Bregni, Achille Pattavina, Gianluca Vegetti
Dept. of Electronics and Information, Politecnico di Milano
{bregni, pattavina}@elet.polimi.it

Abstract— **In order to support the continuous growth of transmission capacity demand, optical packet switching technology is emerging as a strong candidate, promising to allow fast dynamic allocation of WDM channels, combined with a high degree of statistical resource sharing. This work addresses the comparison of input and shared buffering in an optical switch architecture based on an arrayed waveguide grating to route packets. Since a limited maximum buffer depth is feasible nowadays, particular care must be paid to buffers implementation. The performance evaluation considers different length distributions of packets offered to the node.**

I. Introduction

In the latest years telecommunication networks have been demanding an unprecedented, dramatic increase of capacity, fostered mostly by the exponential growth of Internet users and by the introduction of new broadband services. The IP architecture is being seen as the unifying paradigm for a variety of services and for making real the Broadband Integrated Services Digital Network (B-ISDN). To face this challenge, considerable research is currently devoted to design IP fully-optical backbone networks, in order to relieve the capacity bottleneck of classical electronic-switched networks.

A single optical fiber offers a potentially huge transmission capacity: just in the third wavelength window, tens of terahertz are there to be mined, if only we could be able to exploit such tremendous bandwidth with adequate technology. In the last ten years, optical Dense Wavelength Division Multiplexing (DWDM) has been developed, which made available commercial systems providing impressive transmission capacities: one terabit per second per fiber, over distances on the order of 100 km, are feasible nowadays. Moreover, recently DWDM has evolved to support some network functions as circuit routing and wavelength conversion and assignment. In the future, Optical packet switching will provide an efficient bandwidth exploitation ([1], [2], [3]).

We address here the long-term view of a full packet switching network performing IP packet transport, in which optical operations are performed as much as possible exploiting the currently available optical device technology. This paper deals with the comparison between input and shared buffering in the architecture of an optical packet switching node previously proposed in [4] and [5]. The paper is organized as follows. Section II describes the optical network architecture we envision and the proposed architecture of an optical packet switching node. Section III provides the evaluation of traffic performance for the switching node for which three different distributions of packet lengths have been considered.

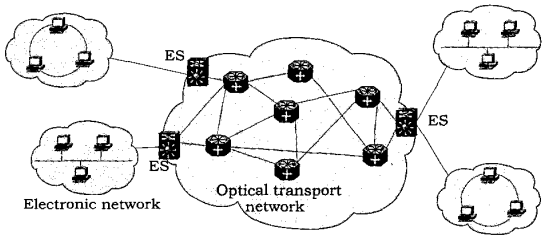

Fig. 1. Optical transport network architecture.

II. Optical network and node architecture

The architecture of the optical network we propose includes $M = 2^m$ optical packet switching nodes, each denoted by an unique address of $m = \log_2 M$ bits. A number of edge systems (ES) interfaces the optical transport network with IP legacy electronic network (as shown in Fig. 1). An ES receives packets from different electronic networks and performs the optical packet generation. The optical packet is composed of a header, containing the m-bit destination address and the payload length indication, while the payload may be composed of a single or, eventually, multiple IP packets with the same destination ES. Optical packet and header are transmitted using SCM (Sub Carrier Multiplexing) technique (see, e.g, [6]) in order to maximize bandwidth utilization by the payload.

The transport network operation is asynchronous, i.e. packets can be received by nodes at any instant and no delineation is needed at the node output. On the other hand, the internal operation of the optical node is synchronous, or *slotted*; in this model we propose the time-slot duration T to be equal to the transmission time of an optical packet whose payload is just the smallest TCP/IP packet, that is the *acknowledgment* packet, whose length is 40 bytes.

Each switching node is fed by N incoming fibers, each carrying W WDM channels. At each switching node, WDM channels are demultiplexed and fed to the structure shown in Fig. 2, one channel per input. This structure consists of three stages: (i) a synchronization stage, as packets are time-slot aligned when entering the node; (ii) a fiber delay line stage, used to delay packets for contention resolution, which acts as an input buffer since a group of delay lines is uniquely

0-7803-7710-9/03/$17.00 © 2003 IEEE.

associated to each node input channel; (iii) a switching unit where packets are routed to the addressed output. At the switching matrix outputs, the WDM channels are multiplexed on outgoing fibers, in order to be transmitted to the next node. The synchronization unit consists of a series of 2×2 optical switches interconnected by fiber delay lines of different lengths. These are arranged in a way that, depending on the particular path set through the switches, packets can be delayed for a variable amount of time, ranging between $\Delta t_{min} = 0$ and $\Delta t_{max} = 2(1 - (1/2)^{n+1}) \times T$, with a resolution of $T/2^n$, where T is the time-slot duration and n the number of delay line stages. The input buffer stage is composed of fiber delay lines able to provide delays ranging from a null delay up to a value of D_{max} seconds, with a resolution of T seconds $(0, T, 2T, 3T, ..., D_{max})$. Packets are routed to the desired delay line by means of a Tunable Wavelength Converter (TWC) (see e.g. [7] and [8]) and a demultiplexer, which routes an input signal to one of its outlets, according to the signal wavelength. The switching stage (see Fig. 3(a)) is realized as a combination of TWCs and an Arrayed Waveguide Grating (AWG), [9]. k different wavelengths are multiplexed on each AWG input port, while R ports of the AWG can be used to implement a shared buffer, using recirculation delay lines, which allow to further delay a packet even when it has already entered the switching stage. Since a minimization of the node hardware complexity is obtained adopting an AWG with size $W \times W$ (as is shown in [10]), the structure parameters are linked by the following relation (see Fig. 3(a)):

$$N\frac{W}{k} + R = W \implies k = \frac{NW}{W - R} \qquad (1)$$

Each recirculation line delays packets by an amount of time equal to D_{rec} seconds and allows the recirculation of up to W packets at a time, exploiting the W different wavelengths being used. A packet can perform up to r_{max} loops in the delay lines before being transmitted. Since long packets could engage more recirculation ports at a time, only packets whose duration is smaller than or equal to D_{rec} are allowed to perform more than one loop. In order to lower hardware complexity (in terms of number of components needed), up to P recirculation ports can be connected to a single recirculation line, thus obviously lowering the shared buffer capacity (see Fig. 3(b)). Contention resolution is performed in two different ways:

1) *Wavelength conversion*, since packets can reach the addressed output fiber from any AWG input port using W/k different wavelengths; in fact these wavelengths enable the packet to reach one of the W/k AWG output ports multiplexed onto the same output fiber.
2) *Packets delay*, which can be performed in the input buffer stage, in the recirculation lines, or in both stages.

Finally, note that the implementation of recirculation lines, could be used to provide different packet priority classes. In fact, once a packet has reached the switching stage, it can be further delayed with recirculation lines, if a higher priority packet enters the node. This could not be performed when employing only input delay line.

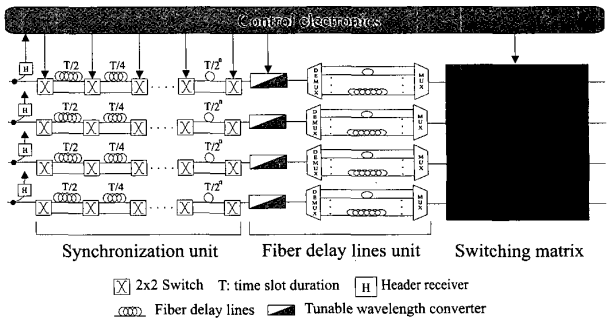

Fig. 2. Switching node architecture.

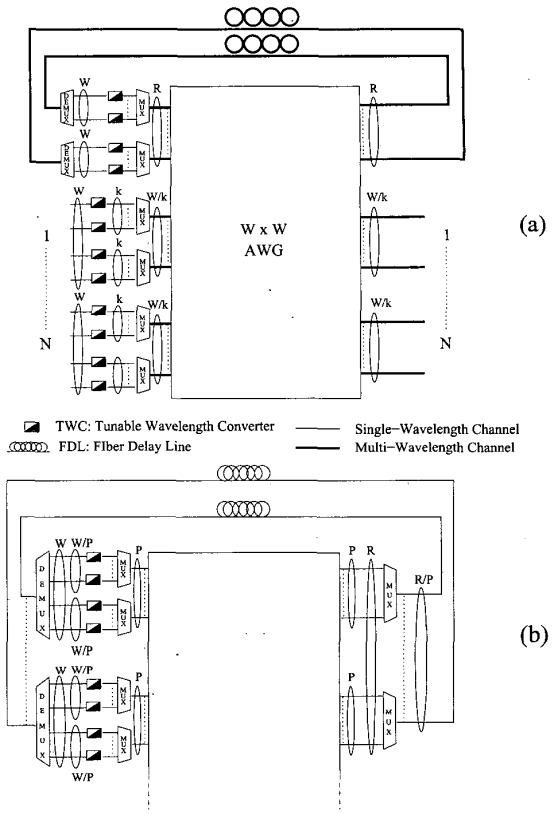

Fig. 3. Switching matrix structure, with $P = 1$ (a) and $P > 1$ (b).

The reader interested in a deeper analysis of each stage is referred to [4], [5] and [10], where an extended structure using more parallel switching planes is described.

III. Simulation Results

In this section, different node architectures are compared in order to investigate on the effectiveness of input and/or shared queueing. The results presented in this section are obtained using computer simulation. The developed simulator keeps running as long as the packet loss probability cannot be evaluated with a 95% confidence level. An interval size

of 5% of the central value is considered for packet loss probabilities greater than 10^{-6}, while a 20% size for lower values.

A. Traffic models

Different traffic models have been used. Packet arrivals have always been modeled as a Poisson process (i.e. exponential distributed interarrivals) for each wavelength. Packet destinations have been supposed to be uniformly distributed over the N output fibers. Packet length is denoted by L (bytes), while packet time duration by L_t (seconds). Three different distributions for the packet length L have been considered:

- *Empirical (or Trimodal) distribution.* Based on real IP traffic measurements [11], the following packet length distribution has been considered:

$$\begin{cases} p_0 = P(L = 40\,\text{bytes}, L_t = T) = 0.6 \\ p_1 = P(L = 576\,\text{bytes}, L_t = 15T) = 0.25 \\ p_2 = P(L = 1500\,\text{bytes}, L_t = 38T) = 0.15 \end{cases}$$

  In this distribution packets have an average length of 393 bytes, equal to a $10T$ time duration. When not differently indicated, this distribution will be used.
- *Uniform distribution.* Packet lengths are equally distributed between 40 bytes and 760 bytes, with a 40 bytes resolution, so that packets time duration is equally distributed between $T, 2T, 3T, ..., 19T$, with an average duration of $10T$.
- *Pareto distribution.* Pareto density function is given by:

$$f_L(x) = \begin{cases} \frac{\alpha k^\alpha}{x^{\alpha+1}} & \text{if } x \geq k \\ 0 & \text{if } x < k \end{cases}$$

  In order to obtain performance results comparable with those given by the empirical distribution, the packet length has been limited to the maximum value of 1500 bytes. We chose to set $\alpha = 1.3$, which is a commonly found value in literature for Pareto traffic modeling, and $k = 19.75$ bytes, in order to obtain 60% of the packets with a duration $L_t \leq T$, as it was in trimodal distribution.

These distributions are shown if Fig 4.

B. Numerical results

First, we will briefly report some results previously presented in [5] and [10], in which input and shared buffer performance were analyzed. Figure 5 reports the packet loss probability for a switching node with $W = 32$, a 32×32 AWG, $D_{max} = 8T$, for different values of node size N. We can see that packet loss probability increases with N. In fact, for greater values of N, the number of contentions that can be resolved in the wavelength domain is reduced, since for every couple of AWG input and node output fiber, only W/N wavelengths can be used, since, from relation (1) $k = N$ when $R = 0$.

Usually, we would expect that a shared buffer would provide a better performance, but in our case such improvement is not granted. In fact, in this architecture the implementation of such

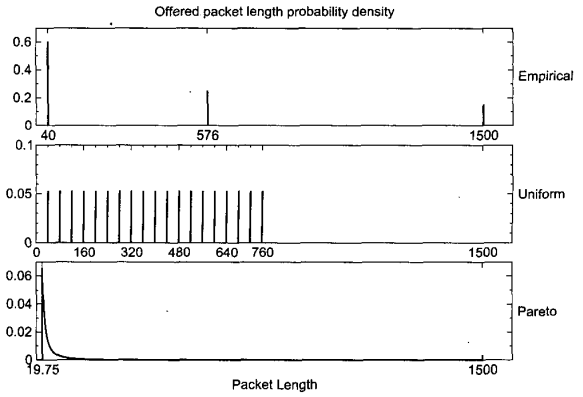

Fig. 4. Packet length distributions.

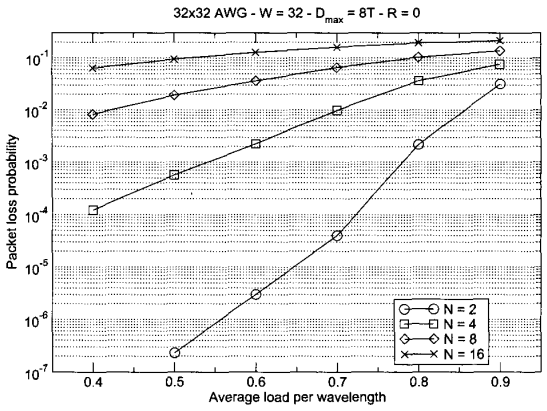

Fig. 5. Packet loss probability for different values of N with $W = 32$.

a buffer using recirculation delay lines reduces the number of AWG ports connected to each output fiber and therefore also the number of contentions resolved in the wavelength domain decreases. Figure 6 shows the packet loss probability of a shared-buffered switching node. A 12×12 AWG is used with $N = 2$ input/output fibers with $W = 12$ wavelengths per fiber and $R = 4$ recirculation ports, the maximum numer of loops per packet has been set to $r_{max} = 10$. We can see how node performance obtained is better than with pure input-buffer in terms of packet loss probability. As for the average packet delay, in Fig. 7 we can see how the increased delay due to shared buffer is relatively limited. In fact, even under heavy traffic conditions, the difference between the average delay in the case of $R = 0$ (where the maximum achievable delay is equal to D_{max}, i.e. $8T$ seconds) and in the case of $R = 4$, $D_{rec} = 16T$ (where the maximum delay is $D_{max} + r_{max} D_{rec}$, i.e. $168T$ seconds) is limited to $8T$ seconds.

Given these considerations, now we compare the traffic performance of a node architecture with either pure input buffering or pure shared buffering. Let us first consider a 12×12 AWG implemented in a switching node with $N = 2$ input/output fibers. We will examine two different structures:

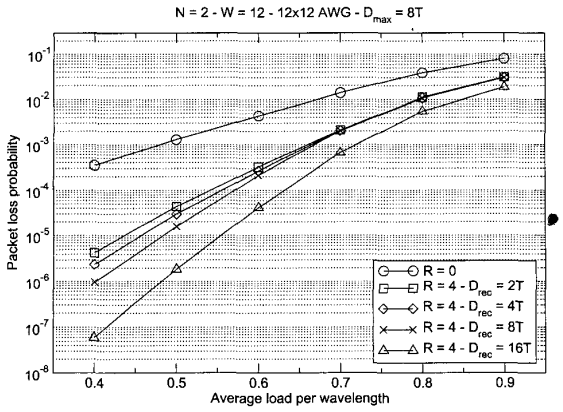

Fig. 6. Packet loss probability for $R = 0$, $R = 4$, for different D_{rec} values.

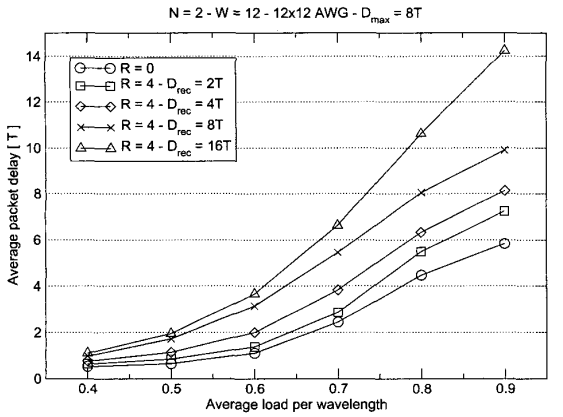

Fig. 7. Average packet delay for $R = 0$, $R = 4$, for different D_{rec} values.

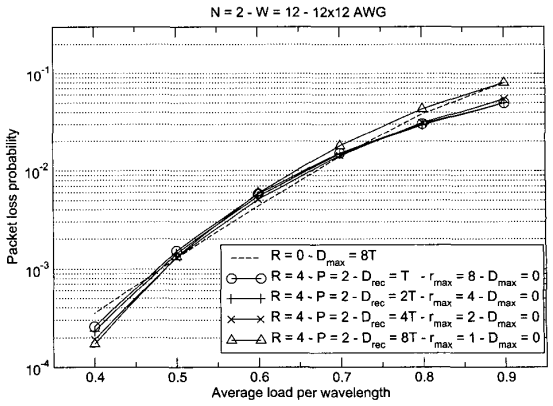

Fig. 8. Packet loss probability for pure input and pure shared-buffered architectures with $8T$ maximum delay.

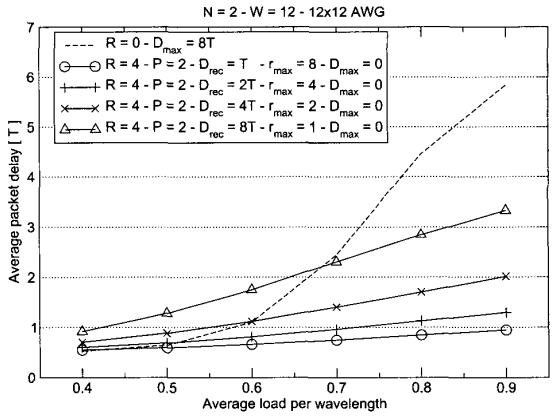

Fig. 9. Average packet delay for pure input and pure shared-buffered architectures with $8T$ maximum delay.

the pure input-buffered structure, where $R = 0$ and $D_{max} = 8T$ and the pure shared-buffered structure, where $R = 4$ and $D_{max} = 0$. In the shared-buffered structure, we will connect $P = 2$ recirculation ports to each recirculation line, in order to obtain a shared buffer which can recirculate up to $W \times R/P = 24$ packets for each time-slot, that is the same capacity of the input buffer, where one packet per input can be buffered ($NW = 24$). In Figs. 8 and 9 four different values of D_{rec} are considered and the maximum number of loops r_{max} a packet can perform is tuned in order to achieve a maximum delay of $8T$, as in the case of input buffering. Therefore, as $D_{rec} = \{T, 2T, 4T, 8T\}$, $r_{max} = \{8, 4, 2, 1\}$. While packet loss probability reported in Fig. 8 is almost equivalent for input and shared buffering, average packet delay, shown in Fig. 9, is lower for shared buffer when the average load is high.

Given this last consideration, we will now examine the same node architecture, where the r_{max} value is no more limited to obtain an $8T$ maximum delay, but is always set to $r_{max} = 10$. In Fig. 10 the packet loss probability of this architecture is compared with that of the input-buffered node, and results in a better performance for heavy and light traffic loads, while for medium traffic the loss probability is almost the same as the input-buffered architecture. Figure 11 shows that the average packet delay given by $r_{max} = 10$ is just a little higher than in previous case.

Let us now evaluate how the packet loss probability is distributed on the different packet lengths. Figure 12 reports the lost packet length distribution for a $\rho = 0.8$ traffic load per wavelength. It can easily be seen that, for the pure input architecture ($R = 0$), the lost packets length follows the offered packet distribution (60% of packets with $L = 40$ bytes, 25% of packets with $L = 576$ bytes and 15% of packets with $L = 1500$ bytes), while for pure shared-buffered architectures the fraction of lost packets is low for 40 byte-packets and consequently increased for the other two packet lengths. Smaller packets are therefore better handled with shared buffer architectures. We can now observe that, as was pointed out in section II, only packets with duration $L_t \leq D_{rec}$ are allowed more than one recirculation. So, as only 40 byte-packets have a duration lower than the maximum D_{rec} value $8T$, only these

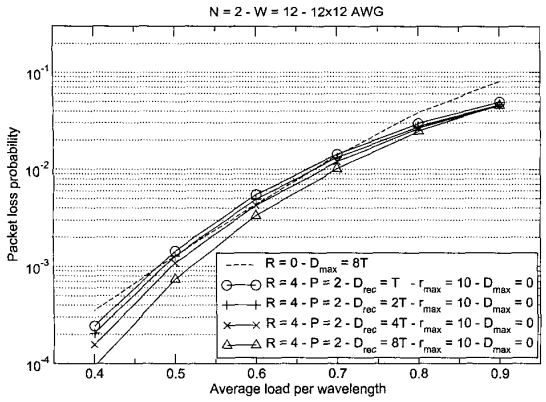

Fig. 10. Packet loss probability for pure input and pure shared-buffered architectures with $r_{max} = 10$.

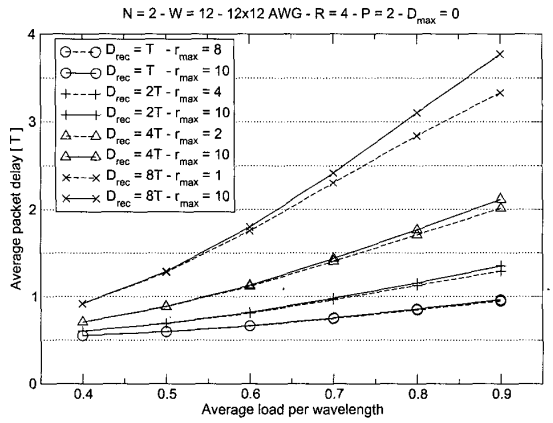

Fig. 11. Average packet delay for pure shared-buffered architectures.

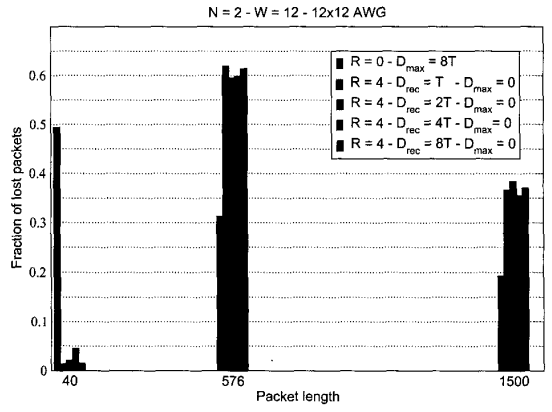

Fig. 12. Lost packet length distribution for pure input and pure shared-buffered nodes with trimodal distribution of offered packet length.

packets can be buffered for longer times than longer packets.

The same analysis is performed in Fig. 13 and in Figs. 14, 15 and 16, where the other different proposed packet

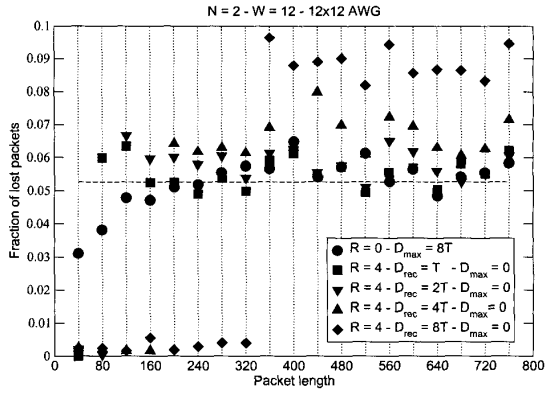

Fig. 13. Lost packet length distribution for pure input and pure shared-buffered nodes with uniform distribution of offered packet length.

length distributions are offered. The uniform distribution is considered in Fig. 13. The horizontal dashed line represents the expected fraction of offered packets for each length value; apparently such value is constant and equal to $1/19$. We can easily see that again, in the case of pure input buffer, the length distribution of lost packets is similar to that of the offered packets, while, in the case of pure shared buffer, the fraction of lost packets is always small when $L_t \leq D_{rec}$ (note that, as 40 bytes are transmitted in T seconds, $L = 40$ bytes means $L_t = T$, $L = 80$ bytes means $L_t = 2T$, etc.). The Pareto distribution is assumed in Figs. 14, 15 and 16. Figure 14 reports the length distribution of lost and offered packets in the case of the pure input-buffered node. The two lines are almost superimposed, and so, once again, the lost packet length distribution follows that of the offered packets. Figures 15 and 16 show the length distribution of lost and offered packets in the shared-buffered node, when $D_{rec} = \{2T, 4T\}$. Also in these cases, only a small fraction of packets with $L_t \leq D_{rec}$ are lost. The performance results given by other values of D_{rec} are similar to those shown here for the two values $2T$ and $4T$, but are not shown for the sake of conciseness. Therefore we conclude that in pure shared-buffered architectures, packet loss probability is smaller for $L_t \leq D_{rec}$.

IV. Conclusions

In this paper the behavior of an optical packet switching node has been evaluated, in order to compare the performance of different solutions, adopting either input or shared buffering. The two architectures interfacing the same number of WDM channels provide similar performance results; only the average delay turns out to be lower with shared buffering, when the offered load is high. Moreover, the implementation of shared buffering leads to lower loss probability for packets with transmission time lower than the recirculation delay. So, in this case, the packet length plays the same role as adopting priority classes for switching packets; in fact, shortest packets are less likely to be dropped than longer ones. Further topics of research could be the real implementation of different

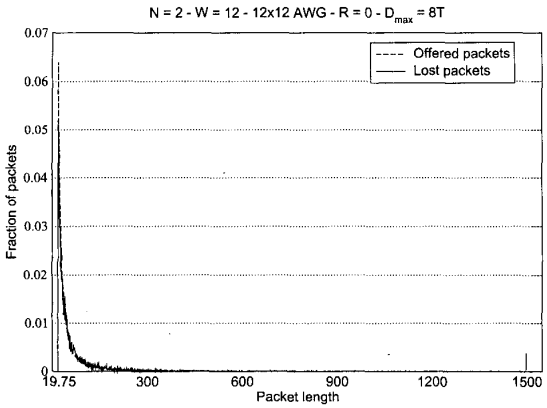

Fig. 14. Lost packet length distribution for pure input-buffered nodes with Pareto distribution of offered packet length.

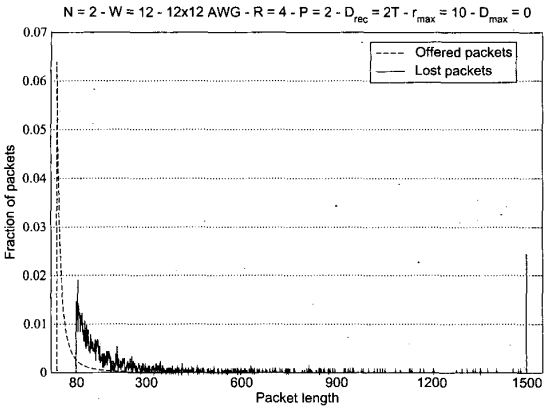

Fig. 15. Lost packet length distribution for pure shared-buffered nodes with $D_{rec} = 2T$ and Pareto distribution of offered packet length.

priority classes, using recirculation lines in the switching node, and a deeper performance analysis using more realistic traffic patterns, such as self-similar traffic traces.

REFERENCES

[1] S. Yao, B. Mukherjee, and S. Dixit, "Advances in Photonic Packet Switching: An Overview," *IEEE Communications Magazine*, vol. 38, no. 2, pp. 84–94, Feb 2000.

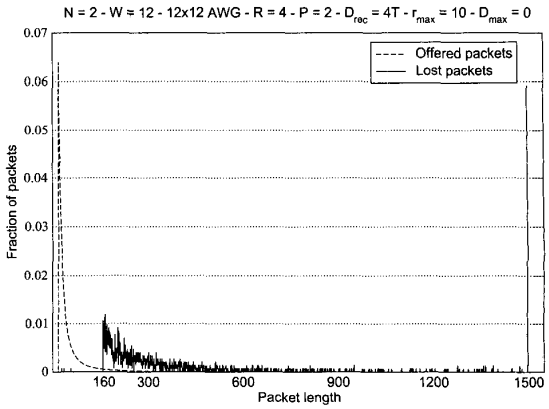

Fig. 16. Lost packet length distribution for pure shared-buffered nodes with $D_{rec} = 4T$ and Pareto distribution of offered packet length.

[2] M. A. Bourouha, M. Bataineh, and M. Guizani, "Advances in Optical Switching and Networking: Past, Present, and Future," in *Proceedings of IEEE SoutheastCon 2002*, 2002, pp. 405–413.

[3] D. K. Hunter and I. Andonovic, "Approaches to Optical Internet Packet Switching," *IEEE Communications Magazine*, vol. 28, no. 9, pp. 116–122, Sep. 2000.

[4] S. Bregni, G. Guerra, and A. Pattavina, "Optical Switching of IP Traffic Using Input Buffered Architectures," *Optical Network Magazine*, vol. 3, no. 6, pp. 20–29, Nov.-Dec. 2002.

[5] S. Bregni, A. Pattavina, and G. Vegetti, "AWG-Based WDM Switching Architectures for All-Optical IP Networks," in *Proceedings of* 7^{th} *Working Conference on Optical Network Design and Modeling (ONDM)*, 2003, pp. 1037–1051.

[6] L. Xu, H. G. Perros, and G. Rouskas, "Techniques for Optical Packet Switching and Optical Burst Switching," *IEEE Communications Magazine*, vol. 39, no. 1, pp. 136–142, Feb. 2001.

[7] A. Tzanakaki and M. O'Mahony, "Analysis of Tunable Wavelength Converters Based on Cross-Gain Modulation in Semiconductor Optical Amplifiers Operating in the Counter Propagating Mode." in *IEE Proceedings-Optoelectronics*, vol. 147, Feb. 2000, pp. 49–55.

[8] M. Mak and H. Tsang, "Polarization-insensitive Widely Tunable Wavelength Converter Using a Single Semiconductor Optical Amplifier." *IEE Electronics Letters*, vol. 36, pp. 152–153, 2000.

[9] C. Parker and S. Walker, "Design of Arrayed-Waveguide Gratings Using Hybrid Fourier-Fresnel Transform Techniques." *IEE Journal on Selected Topics in Quantum Electronics*, vol. 5, pp. 1379–1384, 1999.

[10] S. Bregni, A. Pattavina, and G. Vegetti, "Architecture and Performance of AWG-based Optical Switching Nodes for IP Networks," *To appear in JSAC special issue on High Performance Optical/Electronic Switches/Routers for High Speed Internet*, 2003.

[11] K. Thompson, G. J. Miller, and R. Wilder, "Wide-Area Internet Traffic Patterns and Characteristics," *IEEE Network Magazine*, pp. 10–23, Nov. 1997.