Performance of deflection routing algorithms in IP optical transport networks

Achille Pattavina*

Department of Electronics and Information, Politecnico di Milano, 20133 Milano, Italy

Available online 4 June 2005

Abstract

This paper deals with optical packet switching in a full-IP transport network scenario. For the switching of IP packet flows different node architectures are considered that are based on current optical routing devices. The traffic performance of a mesh network is evaluated with the various node structures, assuming that nodes employ either shortest path routing or deflection routing to forward packets to the addressed destinations. The paper shows how the different node structures behave in terms of packet loss probability with different network configurations when the node parameters are varied.

© 2005 Elsevier B.V. All rights reserved.

Keywords: Optical network; IP traffic; Deflection routing; Arrayed Waveguide Grating (AWG); Packet loss performance

1. Introduction

In the latest years the outbreak of Internet and broadband services for the development of electronic commerce, entertainment and education has involved the increase of demand for transmission bandwidth. Nowadays routing of traffic flows in transport networks occurs by processing electronically data and transmitting them in optical fibers; optics is therefore used exclusively at the physical level. Advent of the wavelength multiplexing technologies (WDM) has concurred an increase of the transmission capacity per fiber but at the same time has shown the limits of actual network infrastructures based on electronic devices. Today networks use only a small fraction of the large-capacity made available by each fiber (in the order of Terabits per second), since electronic switching, processing and storage technologies do not allow to manage fully the huge size of data transported by fibers. At the same time the current need for transporting very large amounts of traffic based on the IP protocol has shown all the limits of SONET/SDH technologies, born as solutions...
for circuit switching and thus unable to guarantee efficient management of the IP traffic flows.

The development of new switching systems is therefore important in order to face such new networking scenario. The advent of optical switching devices is going to define a new generation of network elements in which routing is operated without optical/electronic conversions. In this context optical circuit and packet switching technologies play a different role. With optical circuit switching every connection needs the reservation of an entire WDM channel in order to realize end-to-end circuits, with the inefficient bandwidth usage typical of circuit switching. Optical packet switching on the other hand enables a high and efficient exploitation of the available capacity thanks to the bandwidth sharing typical of statistical multiplexing. This latter technology would moreover be consistent with the new paradigm of an IP protocol that supports any kinds of telecommunication service.

Unfortunately, today optical devices used in market equipment are still too crude to allow packet-by-packet operation. An interesting solution which tries to represent a balance between circuit switching low hardware complexity and packet switching efficient bandwidth utilization is optical burst switching [1,2]. In an optical burst switching system, the basic units of data are bursts, made up of multiple packets, which are sent after control packets, carrying routing information, whose task is to reserve electronically the necessary resources on the intermediate nodes of the transport network.

This paper addresses the future-looking scenario of optical packet switching by exploiting the optical switching technologies available today. In particular it considers the architecture of optical packet switching nodes already proposed in [3–5], which exploit arrayed waveguide grating devices for packet routing and are equipped with fiber delay lines used either for input buffering or for shared buffering of optical packets. Here a further approach is used to resolve contention for the node output links, that is deflection routing. Both shortest-path deflection and random deflection are considered and their traffic performance compared.

The paper is organized as follows. Section 2 describes the envisioned optical network architecture and the routing algorithms adopted, while Section 3 details the proposed structures of the optical packet switching nodes. Section 4 provides a traffic performance comparison of the different node architectures in a mesh network with the different packet routing strategies.

2. Optical network

2.1. Network architecture

The architecture of the optical transport network we propose consists of optical packet-switching nodes, which are mutually connected in a mesh-like topology. A number of edge systems (ES) interfaces the optical transport network with IP legacy (electronic) networks (see Fig. 1). An ES receives packets from different electronic networks (or local hosts) and performs optical packet generation. The optical packet is composed of a simple optical header, which comprises the destination address, and of an optical payload made of a single IP packet, or, alternatively, of an aggregate of IP packets. The optical packets are then buffered and routed through the optical transport network to reach their destination ES, which delivers the traffic it receives to its intended destination in the

Fig. 1. The optical transport network architecture.
electronic domain. At each intermediate node in the transport network, packet headers are received and electronically processed, in order to provide routing information to the control electronics, which will properly configure the node resources to switch packet payloads directly in the optical domain.

The transport network operation is asynchronous; that is, packets can be received by nodes at any instant, with no time alignment. The internal operation of the optical nodes, on the other hand, is assumed to be synchronous, or slotted, since the switching of packets in an unslotted node is less regulated and more unpredictable, resulting in a larger contention probability. In our model we propose, the time slot duration, $T$, to be equal to the amount of time needed to transmit from an input WDM channel to an output WDM channel an optical packet with a 40-bytes payload, which corresponds to the smallest payload enabled by an IP packet. Supposing a bit rate of 10 Gbit/s per wavelength channel, a 40 ns slot duration has been selected; in fact it is assumed that the synchronization/switching operations needed in the node are carried out using semiconductor optical amplifier (SOA) technology, whose switching time is in the nanosecond range. These latter operations together with the optical packet header recovery and processing account for up to 8 ns, so that 32 ns are left for the transmission of a minimum-size IP packet. For a deeper discussion about this issue the reader is referred to [5].

In order to relieve the complexity of the problem related to the arbitrary mesh topology of the transport network, we assume here that all nodes of the transport network have the same nodal degree $N_h$, that is the same number of adjacent nodes. Since we have selected $M = 8$ optical nodes in the transport network, we are going to examine here the four network topologies represented in Fig. 2, where the number $N_h$ of adjacent nodes is $N_h = 2, 3, 5, 7$. Hence these regular topologies span from the ring network ($N_h = 2$) up to the full-mesh network ($N_h = 7$). The other two networks identify topologies with intermediate meshing degree. If we define the connectivity degree $\alpha$ as the ratio between the number of links in the network and that of a full-mesh network, it follows that the eight-node networks in Fig. 2 are characterized by a connectivity degree $\alpha = 0.29, 0.43, 0.71, 1.00$.

2.2. Routing algorithms

Routing of optical packets, which is decided by the control electronics, is carried out based on the shortest path routing concept. Therefore routing tables are held indicating for each received packet the outgoing optical link from the switch to be selected for reaching a given optical node that interfaces the intended destination ES. This paper investigates the issue of adopting deflection routing to solve contention resolution for the intended output link in addition to the other means already available for this scope, that is buffering (either input or shared) and wavelength conversion.

Two different approaches are considered here for accomplishing deflection routing, that is shortest path deflection routing and uniform deflection routing. With shortest path deflection each node keeps a table that for every destination optical node ranks the different paths to be consecutively scanned in order to path lengths somehow measured (for example in terms of number of crossed

![Fig. 2. Network topologies.](image-url)
The output link from the node is selected as the first available in the rank provided. With uniform deflection instead, upon unavailability of the link along the shortest path, a different inter-node available output link is selected randomly. This policy is analogous to “hot potato routing” in classical packet networks. Under uniform deflection, if the node is equipped also with shared buffering, packet recirculation in the shared buffer is exploited only if all the inter-node outgoing links are busy.

3. Node architecture

The general architecture of a network node is shown in Fig. 3. It consists of three stages: a first stage of channel demultiplexing, a second stage of switching and a third stage of channel multiplexing. The node is fed by \(N\) incoming fibers each having \(W\) wavelengths. In the first stage the incoming fiber signals are demultiplexed and \(G\) wavelengths from each input fiber are fed into each one of the \(N_p = W/G\) second-stage switching planes, which constitute the switching fabric core. Once signals have been switched in one of the parallel planes, packets can reach every output port through multiplexing carried out in the third stage using any of the \(G\) wavelengths that are directed to each output fiber. We note that the number of inlets of each third-stage multiplexer varies, depending on the specific structure of the switching planes. Wavelength conversion must be used for contention resolution, since at most \(G\) packets can be concurrently transmitted by each second-stage plane on the same output link.

The detailed structure of one of the \(W/G\) parallel switching planes is presented in Fig. 4. It consists of three main blocks: an input synchronization unit, as the packet switching is slotted and incoming packets need to be slot-aligned, a fiber delay lines unit, used to store packets for contention resolution, and a switching matrix unit, adopted to achieve the switching of signals.

These three blocks are all managed by an electronic control unit which carries out the following tasks:

- optical packet header recovery and processing;
- managing the synchronization unit in order to properly set the correct path through the synchronizer for each incoming packet;
- managing the tunable wavelength converters inside the switching matrix, in order to properly delay and route incoming packets.

The details of the synchronization and of the fiber delay unit can be found in [5]. We simply recall here that the delay lines are used as an optical scheduler that, thanks to the electronic control, maximizes the number of packets transmitted on the requested output links. Given the maximum achievable delay \(D_{\text{max}}\) slot, for each switch input \(D_{\text{max}} + 1\) delay lines are needed in each plane, with
delays growing from 0 to $D_{\text{max}}$. Once packets have crossed the fiber delay lines unit, they enter the switching matrix stage in order to be routed to the desired output port. This is achieved using a set of tunable wavelength converters combined with an arrayed waveguide grating (AWG) wavelength router [6].

The AWG is used here as it gives better performance than a normal space switch interconnection network, as far as insertion losses are concerned. This is due to the high insertion losses of all the high-speed all-optical switching fabrics available at the moment, that could be used to build a space switch interconnection network. Commercially available 40 channel devices have a channel spacing of 100 GHz and show an insertion loss of less than 7.5 dB [7]. Other proposals of switching nodes based on AWGs can be found in [8,9].

Three different structures are proposed here for the implementation of the switching matrix unit, referred to as structures Basic (B), Enhanced (E) and Optimized (O). In all these structures a shared buffer is implemented through recirculation lines in order to enable a much more effective contention resolution. $R$ denotes the number of AWG ports destined to recirculation lines, each one delaying the packets by a fixed amount, $D_{\text{rec}}$ (slot), and $R_{\text{max}}$ denotes that maximum number of recirculations allowed to a packet in the node.

3.1. Basic structure (B)

The simplest switching matrix (Basic structure), first proposed in [3], is shown in Fig. 5, referred to a single plane. It consists of $2NG + R$ tunable wavelength converters and an AWG with size $(NG + R) \times (NG + R)$. Only one packet is routed to each AWG outlet and this packet must finally be converted to one of the wavelengths used in the WDM channel, paying attention to avoid contention with other packets of the same channel.

3.2. Enhanced structure (E)

In order to reduce the number of planes of the node and thus to better exploit the “channel grouping” effect (i.e. the sharing of different channels for transmitting a large number of packets, the load per channel being constant), more than one packet can be routed through each AWG inlet; apparently the packets sharing the same input must be transmitted on different wavelengths. The structure of the AWG is such that different wavelengths entering the same input port will emerge on different output ports.

In the Enhanced switching matrix structure illustrated in Fig. 6, up to $k$ different packets are sent to the same AWG inlet using different wavelengths. A simple node design requires $k$ to be an
integer that divides $G$. From AWG input port $i$, the output channel $j$ can be reached by $Gk$ different packets, since there are exactly $Gk$ AWG outlets connected to that channel. During each time slot, up to $G$ packets can be routed to the same AWG outlet using different wavelengths. Hence, demultiplexers are needed to split the different signals and to route them to the last stage of wavelength converters. If $k \leq Gk$, no contention can happen in the multiplexing stage, so this structure behaves exactly as a structure Basic with size $NG\times NG$. On the other hand, when $k > Gk$, events of packet blocking occur, considering the fact that $Gk$ paths are available to reach a tagged output for up to $k$ packets per inlet. So, when more than $Gk$ packets in the same AWG inlet are destined to the same output channel, a contention happens, even if the total number of packets addressed to that output is smaller than $G$.

### 3.3. Optimized structure (O)

The Enhanced node structure be simplified by selecting $k = N$, so that each AWG input can receive up to $N$ packets using different wavelengths, thus implying that the node structure includes just one plane ($N_p = 1$). Therefore, the number of AWG inlets is now exactly $W$. In this structure, shown in Fig. 7 and referred to as Optimized structure, the last TWC stage is not needed anymore, provided the employed AWG works on the same wavelengths used in the outgoing fibers. In fact, if the electronic controller takes care of avoiding wavelength contention between AWG outlets connected to the same output channel, packets are ready to be transmitted as soon as they exit the AWG. Therefore, a packet entering the AWG inlet $i$ and destined to the output WDM channel $j$ can not be transmitted using every color in the WDM channel, but only using a subset which consists of the $W/N$ wavelengths through which the packet can reach the desired output channel, thus reducing the benefits of channel grouping (when $N$ and $W$ are kept constant).

### 4. Performance evaluation

We show now some traffic performance results given by the different node architecture configurations obtained through computer simulation. In order to evaluate the performance provided by the three different switching structures, we will first examine the packet loss probability in 8-node networks with different connectivity factors. Then we will examine the effect of varying one single network parameter, that is the number of nodes, the network connectivity degree $x$ and the AWG size. All the performance figures plotted in this section and obtained through computer simulation are characterized by a 95% confidence interval not larger than 30% of the plotted value.

In order to obtain performance results mutually comparable in spite of the different connectivity degrees, we assume that each network node interfaces a number $N_h$ of edge systems (ES) each generating the same amount of traffic $\rho$. In such a way the total number of ESs interfacing a node is equal to the number of network links outgoing from the node and the total (normalized) load offered to the network is $\rho$. Therefore the switch size $N$ is given by $N = 2N_h$.

Traffic levels and shared buffer capacities are assumed equal for all nodes, so that Basic and Enhanced nodes require more planes ($N_p > 1$). The AWG size is set equal in all nodes and this value depends on the network connectivity. Given the

![Fig. 7. Optimized switching matrix.](image)
wheel network topologies considered here and the aim of assuming reasonable AWG sizes, we have selected as a basic point that the parameter $G = 2$ for the Basic node with shared buffer; furthermore the number of recirculation lines $R$, when equipped, equals half of the AWG size (if $R = 0$ only input buffering is exploited by the node). Given the previous assumption about the ESs (the internode links equal the number of local ESs), it follows that we are assuming an AWG size $8N_h \times 8N_h$; in fact the local interfaces to ESs account for $N_h$ and $G = 2$ wavelengths feed each switching plane from any input link. Then it follows that the AWG sizes are $16 \times 16$, $24 \times 24$, $40 \times 40$ and $56 \times 56$ for the nodal degrees $N_h = 2$, $3$, $5$, $7$, respectively. Furthermore the number of wavelengths $W$ equals the AWG size, since the Optimized node includes just one plane.

The buffering capacity of each node has been set according to the following parameters: $D_{\text{max}} = 8$ slot in input buffers, while $D_{\text{rec}} = 2$ slot and $R_{\text{max}} = 4$ when shared buffering is equipped ($R > 0$).

As far as the offered traffic distribution is concerned, packet interarrivals for each ES wavelength has been modelled as a Poisson process with negative exponential distribution. Based on measurement of real IP traffic [10], the following distribution of packet length $L$ has been assumed

\[
\begin{align*}
    p_0 &= \Pr(L = 40 \text{ bytes}) = 0.6 \\
    p_1 &= \Pr(L = 576 \text{ bytes}) = 0.25 \\
    p_2 &= \Pr(L = 1500 \text{ bytes}) = 0.15
\end{align*}
\]

so that the resulting average packet length is 393 bytes. Packets are assumed to be equally likely to be addressed to any destination ES.

4.1. Shortest path deflection routing

The packet loss probability with shortest path deflection is shown in Figs. 8–11, for $\alpha = 0.29$, $0.43$, $0.71$, $1.00$, respectively. We notice that, as we might expect, solutions with shared buffering give better performance than without it. The Optimized node gives the worst performance without shared buffers, the other two solutions providing similar behaviour. An explanation of this is that packets coming from a local ES or a node cannot be routed to the same node outlet (ES or node). So this implies that with the Enhanced and Optimized solutions more contentions arise for accessing each single node outlet compared to a case in which all node outlets are equally addressable, as assumed in [5].

On the other hand when shared buffering is employed, the Enhanced and Optimized structures outperform the Basic one especially for large network connectivity degrees. This is due to the fact that the shared buffer is distributed among a much larger number of planes in the former node.
structures and then becomes less efficient than a larger buffer in a single plane (Optimized solution).

4.2. Uniform deflection routing

The three switching node structures are now compared with uniform deflection routing for the same values of the network connectivity and the corresponding results are shown in Figs. 12–15.

It is quite interesting to note that, unlike shortest-path deflection, now the network performance improves significantly as the network connectivity grows also with pure input queueing. This can be explained considering that random deflection can be exploited better with more output links from the node. When shared buffering is employed Enhanced and Optimized node structures give again the best performance due to the larger buffers they can exploit, as with shortest-path deflection. Without shared buffering, the Enhanced node gives the best performance, whereas the Optimized node behaves the worst. The reason is analogous to that given for shortest-path routing. In fact the load offered to the node outputs is not evenly distributed, since now packets entering the node from local outlets cannot exit the node on the same outlet.

Fig. 10. Loss performance with shortest path deflection and $\alpha = 0.71$.

Fig. 11. Loss performance with shortest path deflection and $\alpha = 1.00$.

Fig. 12. Loss performance with uniform deflection routing and $\alpha = 0.29$.

Fig. 13. Loss performance with uniform deflection routing and $\alpha = 0.43$. 
Note that now, due to uniform deflection, packets entering the node from an upstream node can be routed on any output. Hence it follows that the load offered to inter-node outlets is larger than that received from local outlets of the node. Notice that such behavior applies to all node structures.

Unlike networks based on Enhanced and Optimized nodes, adding shared buffering with the Basic solution provides worse loss performance. As already observed in Ref. [5], shared buffering with Basic node is beneficial only if the (fixed) delay of recirculation lines exceeds a given threshold, for example 16 slots (recall that we have assumed here a delay $D_{\text{rec}} = 2$ slots).

4.3. Comparison between deflection routing algorithms

Given the results presented insofar for each of the two deflection routing algorithms, we compare now their loss performance directly considering the node architecture that behaves better than the others, that is the extended structure (E). The previous curves for the four network topologies are now shown in Figs. 16 and 17 for the two cases of absence ($R = 0$) and presence ($R > 0$) of shared buffering, respectively. Note that the AWG size

Fig. 14. Loss performance with uniform deflection routing and $\alpha = 0.71$.

Fig. 15. Loss performance with uniform deflection routing and $\alpha = 1.00$.

Fig. 16. Loss comparison for different deflection routing strategies without shared buffering.

Fig. 17. Loss comparison for different deflection routing strategies with shared buffering.
differs from one topology to the other as it depends on the network connectivity; the actual values are the same already described at the beginning of Section 4. In Fig. 16 we note that uniform deflection outperforms shortest path deflection under low levels of offered load. On the other hand a little gain in loss performance is attained by the latter type of deflection when the loss probability is larger than $10^{-2}$. Similar results are observed when shared buffer is adopted in addition to input queueing (see Fig. 17). By comparing the last two plots, we remark that shared queueing is actually able to improve the performance of meshed networks, especially for increasing connectivity degrees.

4.4. Effect of node number, connectivity degree and AWG size

Under the assumption of uniform deflection and absence of shared buffering ($R = 0$) we examine now the effect on network performance of three parameters: the number of nodes $M$, the connectivity degree $\alpha$ and the AWG size.

For an AWG of size $40 \times 40$ and connectivity $\alpha = 1.00$ (full mesh network), three different values of the number of nodes have been considered, that is $M = 3, 5, 6$. Fig. 18 shows that performance of the Basic node degrades as the number of nodes grows; in fact this implies that the number of wavelength channels per plane ($G$) diminishes and hence less benefit can be gained by the channel grouping effect. Analogous behavior is observed with the Enhanced and Optimized structures, since the number of wavelengths multiplexed on the same AWG inlet increases with number of nodes thus determining more contentions.

Fig. 19 shows the network performance for an 8-node network with different connectivity degrees $\alpha = 0.71, 0.43, 0.29$ when the AWG has size $60 \times 60$. As in the previous case, performance improves for higher connectivity degrees, which is explained by the same reason as before. Additionally we have to notice that a small connectivity degree can penalize the network performance under uniform deflection, since a deflection can imply a much longer path to the intended destination, as already observed in Ref. [11].

Different AWG sizes are now considered, that is $14 \times 14$, $28 \times 28$, $42 \times 42$ and $56 \times 56$, for a full-mesh network ($\alpha = 1.00$) with $M = 8$ nodes. In this case, since the AWG size of the Optimized node provides the wavelength number $W$, the wavelength channels per plane in the Basic node associated to an input/output fiber are 1, 2, 3, 4, respectively. As far as the Enhanced node is concerned the number of planes has been selected to $N_p = 2$, so that the number of wavelengths $k$ multiplexed in each AWG inlet is 7. Simulation results are shown in Fig. 20 for Basic and Enhanced node (performance of Optimized node is no shown as it
behaves worse than the other two nodes for all load values. The figure shows that increasing the AWG size has the same beneficial effect as increasing the number of nodes. In fact with the Basic node the channel grouping effect is enhanced as $G$ increases, whereas in the Enhanced structure more AWG ports interface a single output fiber since $k$ is kept constant, thus reducing packet contentions.

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{network_performance.png}
\caption{Network performance with nodes Basic and Enhanced adopting different AWG sizes.}
\end{figure}

5. Conclusions

In this work we have analyzed and compared the packet loss performance of three types of optical nodes that differ substantially for the switching matrix when deflection routing is adopted. We have found that the Enhanced node performs always better than the others without or with shared buffering. In the former case the Basic node provides better results than the Optimized node, whereas in the latter case the situation is reversed. Adding shared buffering to a node by keeping the same AWG size (and hence increasing the number of planes with Basic solution or increasing the multiplexing factor in Enhanced and Optimized solutions) provides large benefits with all nodes except in the case of uniform deflection routing and Basic structure. Adopting uniform deflection improves significantly the loss performance especially for low traffic loads and high connectivity degrees.

Acknowledgment

The author is grateful to Mr. Maurizio Aste for his work on the software package used to get the simulation data reported here, which are part of his Master degree thesis.

References

Achille Pattavina received the degree in Electronic Engineering (Dr. Eng. degree) from University “La Sapienza” of Rome (Italy) in 1977. He was with the same University until 1991 when he moved to “Politecnico di Milano”, Milan (Italy), where he is now Full Professor. He has been author of more than 100 papers in the area of Communications Networks published in international journals and conference proceedings. He has been author of the book *Switching Theory, Architectures and Performance in Broadband ATM Networks* (John-Wiley & Sons). He has been Editor for Switching Architecture Performance of the IEEE Transactions on Communications since 1994 and Editor-in-Chief of the *European Transactions on Telecommunications* since 2001. He is a Senior Member of the IEEE Communications Society. His current research interests are in the area of optical networks and traffic modeling.