Service category-to-wavelength selection technique for QoS support in connection-oriented optical packet switching

D. Careglio *, J. Solé Pareta, S. Spadaro

Advanced Broadband Communications Centre, Universitat Politècnica de Catalunya, Jordi Girona 1-3, 08034, Barcelona, Spain

Abstract

This paper considers an optical packet-switched node subject to asynchronous, variable-length packets and connection-oriented operation. We firstly address the problem of setting up the optical virtual connections and properly configuring the forwarding table at the node. We do not deal with routing aspects, but with the efficient mapping of the virtual connections to the wavelengths of the output ports. In this context, we suggest a wavelength assignment procedure that improves the node performance in comparison with simple random or balanced schemes. We then address the QoS provisioning problem. While existing solutions focus on applying some forms of resource reservation on top of the contention resolution algorithm, here we propose a method based on the well-known ATM scheme of defining different service categories. In particular, we define a case study with three OPS service categories, and for each category a specific contention resolution algorithm is applied. With such a strategy the algorithms present the problem of a different performance alignment; we solve it by designing an ad-hoc optical buffer architecture based on non-degenerate delays. The performance of the final node architecture is evaluated by simulation. The results obtained indicate the merits of this method, which opens up interesting future developments for a whole optical network scenario.

Key words: Quality of service, category of service, connection-oriented optical packet switching, wavelength assignment procedure, performance evaluation
PACS:

* Corresponding author.
Email addresses: careglio@ac.upc.edu (D. Careglio), pareta@ac.upc.edu (J. Solé Pareta), spadaro@tsc.upc.edu (S. Spadaro).

1 Introduction

In recent years packet switching approaches have been gaining the confidence of experts as potential solutions for the next generation high-capacity Internet [1][2]. Instead of installing over-provisioned circuits such as ASON/GMPLS solutions, in the perspective of network optimization the packet switching techniques applied directly in the transport network take advantage of statistical multiplexing. If switching is performed optically (i.e. data remains in the optical domain during the entire source-destination path), the concept is referred to as optical packet switching (OPS) [3].

Packet contention is the main problem in packet switching technologies. A contention resolution policy must be applied to reduce the packet losses and make the statistical multiplexing more efficient. Contention resolution techniques typical exploit the time domain by means of buffering. In OPS, the lack of optical RAMs imposes the use of a pool of Fiber Delay Lines (FDLs), which are bulky, unscalable and offer very limited buffering capabilities [4]. The use of WDM links and wavelength converters becomes the key factor in OPS because it allows packet contention to be solved also in the frequency domain by means of wavelength multiplexing.

Nonetheless, optics is still in its infancy and there are currently some technological constraints in achieving OPS which lie in the following facts. Optical processing is not available, so each node must extract the header of each packet, and convert and process it electrically; meanwhile, the rest of the packet (the payload) can remain in the optical domain, delayed at the input interfaces by means of an FDL, if necessary, to give sufficient time for the processing and switching tasks [3]. Furthermore, OPS requires the use of high-speed optical devices such as wavelength converters and very fast optical switches, which are not yet mature. However, OPS is expected to be deployed in the long-term future when the aforementioned components will probably be commercially available at a cost comparable to that of electrical ones [4].

In this paper we therefore assume the availability of the OPS technology and focus on a generic non-blocking OPS node acting as an output queuing switch with full wavelength conversion capabilities and an optical buffer made by B FDLs. We consider a feed-forward buffer configuration [5], but the concepts developed in this paper are also valid for a feedback configuration (i.e., recirculation buffering). We assume that the node is capable of switching asynchronous, variable-length packets [6], allowing a better interworking with heterogeneous client traffic [7]. The electronic control unit takes all the decisions regarding the configuration of the hardware to perform the proper switching actions.

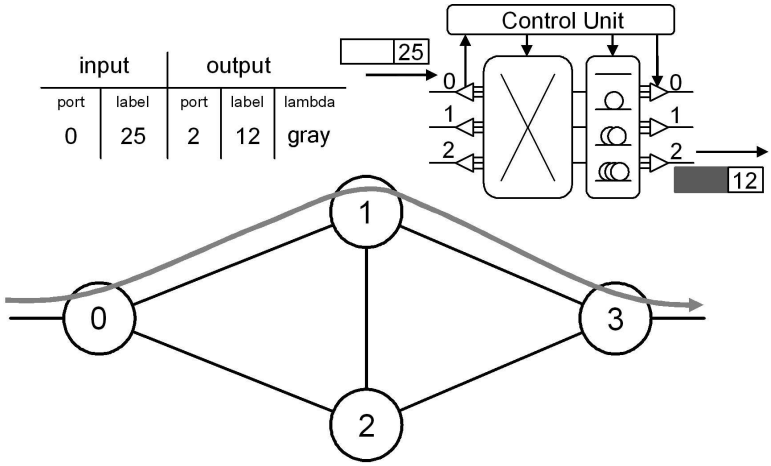

Fig. 1. Connection-oriented OPS network

In order to reduce the control complexity and improve the network performance, recent works [8][9][10] suggest the integration of a *connection-oriented* path management protocol (for instance MPLS) on top of the contention resolution algorithm.

In this MPLS over OPS network, the tasks of the nodes are two-fold. On the connection side, the edge nodes are in charge of setting up and maintaining the unidirectional Label Switched Paths (LSPs) throughout the network by means of a *signaling protocol* such as RSVP, as well as configuring the LSP forwarding table at each core node of the path by means of a *Routing and Wavelength Assignment* (RWA) scheme. On the packet side, based on the destination address and the quality of service requirements, the packets coming from the client networks are classified at the edge nodes into a finite number of subsets such as the *Forwarding Equivalent Classes* (FECs) concept defined in the MPLS environment. Each FEC is identified by an additional *label* added to the packets. As packets belonging to the same FEC are identical from a forwarding point of view, they are transferred from source to destination along the LSP which corresponds to their label. This approach simplifies the tasks of the core nodes; in fact a simple label matching operation on the LSP forwarding table is required for each incoming packet, which speeds up the forwarding function compared with the connectionless OPS case [9] and implicitly reflects the QoS requirements of the packets belonging to their LSPs. Figure 1 shows an example; an LSP forwarding table is setup in node 1 which indicates that the packet with label 25 coming from port 0 should be forwarded to the output port 1 with a gray wavelength and an output label 12.

This approach is similar to the ASON/GMPLS circuit switching solutions, but, as mentioned above, it achieves statistical multiplexing by sharing the wavelengths among several LSPs. To solve possible packet contentions, a contention resolution algorithm must be applied. The use of the connection-oriented approach allows a suitable contention resolution algorithm to be de-

signed which, combining the time and wavelength domain, can fully exploit the packets correlation belonging to the same LSP. A generic algorithm can be schematized as follows:

(1) Lookup the forwarding table to determine the output port n^{out} (also determining the network path), the output wavelength λ^{out}, and the output label l^{out};
(2) If λ^{out} is fully congested;
  (a) Search for the set of wavelength $\Lambda \in n^{out}$ not busy;
  (b) If $\Lambda = \emptyset$, then the packet is lost;
  (c) If $\Lambda \neq \emptyset$, select a new wavelength $\lambda^{out}_{new} \in \Lambda$;
(3) Determine the smallest delay D_j for λ^{out} and select the FDL j;
(4) Transmit the packet to FDL j with wavelength λ^{out}.

While the output port of the node is determined by the routing algorithm, the selections of the wavelength in Step 2 and of the delay in Step 3 are the key points of the contention resolution algorithm. In our scenario they are actually correlated: since each wavelength has its own logical output buffer, choosing a particular wavelength is equivalent to assigning the packet one of the available delays on the corresponding buffer. Here we assume that, once the wavelength has been chosen, the smallest delay available after the last queued packet on the corresponding buffer is always assigned (i.e., we do not consider void-filling algorithms [11]). The smallest delay available on a given wavelength can easily be computed using the smallest integer greater than or equal to the difference between the time when the wavelength will be available again and the packet arrival time.

The step of wavelength selection (and therefore the correlated delay) can be implemented by following two different policies:

- **Static wavelength selection**. The wavelength assigned at the LSP setup is held over the LSP life (i.e., Step 2 is bypassed). Therefore, packets belonging to the same LSP are always switched to the same wavelength and the contentions can be only solved in time (Step 3).
- **Dynamic wavelength selection**. The wavelength assigned at the LSP setup can be changed during the LSP life. When heavy congestion arises on the assigned wavelength (i.e., when the time domain is not able to solve a contention), the LSP is switched to another wavelength and the forwarding table is updated concordantly. Here two alternatives are possible:
  · the new assignment can be kept till congestion arises also on the new wavelength or
  · it can be temporary and the LSP is switched back to the original wavelength when congestion disappears.

The static policy requires minimum control complexity since processing is

performed only at LSP setup. It also preserves the correct order of packets belonging to the same LSP since new arrivals cannot overtake older packets. However, it does not optimize the resources and obtains high packet loss rate figures [8]. On the other hand, when a dynamic policy is executed, the LSP is switched to an alternative wavelength that is not (or is less) congested. As a consequence new incoming packets on that LSP will in general experience less queuing time than older packets and will very likely overtake them along the network path causing an out-of-sequence delivery. Furthermore, the amount of executions of the algorithm affects the processing load of the control unit, ranging from no efforts if a static approach is used to fairly demanding efforts if a new wavelength selection is executed for each incoming packet (e.g., [7] [12]). Therefore, the complexity of the contention resolution algorithm not only depends on its computational complexity but also on the number of times it is executed (i.e., the number of times the wavelength search is executed). This feature will be considered later on when the algorithms are designed and the performance measures are discussed.

In the context of the MPLS over OPS scenario, we address two problems: the problem of setting up the optical LSP and properly configuring the forwarding table at the nodes, and the problem of providing QoS.

Concerning the first problem, at the LSP setup each node must assign both the output port and the output wavelength to the LSP in such a way that the packets belonging to that LSP are always switched to the same output. In comparison with the *classical* RWA problem in circuit-switched network, here the wavelengths are shared among several LSPs (in a packet-switched basis). In this study we do not deal with the problem of assigning the output port, which depends on the routing protocol, but we are interested in the assignment of the wavelength, which may be set locally by each node using a *Wavelength Assignment* (WA) algorithm (see for instance [13]). In particular we show that intelligent WA procedures can considerably improve the performance of the OPS node. The intelligence relies on grouping the LSPs coming from the same input wavelength, which allows us to obtain contention-free situations.

Concerning the second problem, existing solutions to provide QoS in OPS networks are based on the following strategy: 1) design a contention resolution algorithm which minimizes the Packet Loss Rate (PLR), and 2) apply a QoS mechanism (some form of resources reservation on top of the contention resolution algorithm) able to differentiate the PLR between two or more classes. Given that we are dealing with a connection-oriented model, we suggest a new method based on the well-known ATM scheme of defining different service categories. It consists of defining different OPS service categories, each one based on a different contention resolution algorithm specifically designed to cope with the requirements of that category. With this technique, besides the PLR, the packet delay and the overload complexity can also be considered

as QoS metrics.

The rest of the paper is organized as follows. In Section 2 we focus on the LSP setup problem in OPS, proposing and evaluating an efficient WA procedure. In Section 3 we describe the new method based on defining service categories and evaluates its performance by simulation. Section 4 concludes the paper. The simulation environment used in the performance evaluation included in Sections 2 and 3 is described in Appendix A.

2 The WA problem in connection-oriented OPS

2.1 Problem description

In an OPS connection-oriented scenario, the configuration of the LSP forwarding table plays a significant role in improving the network performance. In this context, a basic observation is that packets following LSPs incoming on the same input wavelength do not overlap because of the serial nature of the wavelength as a transmission line. Therefore, such packets contend for output resources only with packets incoming on different wavelengths. As a consequence, as stated in [8], if the LSPs incoming on the same input wavelength are the only ones forwarded to the same output wavelength, contention will never arise (we called it *contention-free* configuration).

Figure 2 shows an example of a node with $N = 2$ ports and $W = 3$ wavelengths per port. On wavelength λ_2^{in} of port n_1^{in} three LSPs are active: two (l_1 and l_2) are switched to λ_1^{out} of n_1^{out}, and the other (l_3) to λ_3^{out} of n_1^{out}. On port n_2^{in} there are three LSPs (l_4, l_5 and l_6) coming from different wavelengths. LSP l_4 is switched to λ_3^{out} of n_1^{out} while l_5 and l_6 are switched to λ_2^{out} of n_2^{out}. By observing the figure it is easy to understand that packets from l_1 and l_2 will never overlap (there are subject to a contention-free configuration), whilst packets from l_3 and l_5 may overlap with packets from l_4 and l_6 respectively.

It is possible to quantify the influence of the overlapping to the node performance. For the evaluation we use the simulation environment described in Appendix A, where we also explain the meaning of the parameters used in this section. The following results were obtained using a node with $N = 4$, $W = 16$ and a degenerate buffer of length $B = 6$. The granularity of the FDL was set to $D = 0.4$ because it is the optimal value for the static wavelength policy [9]. Each input and output wavelength is assumed to carry $L = 10$ different LSPs for a total of 640 incoming LSPs. We concentrated our attention on a particular output wavelength. At different values of offered load ρ, we carried out a set of simulations changing the percentage of the LSPs coming from the

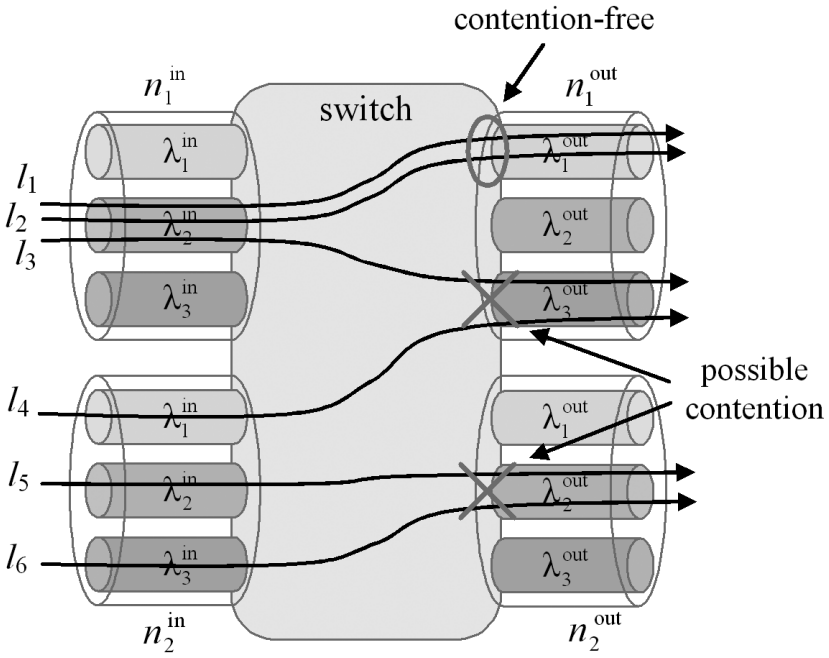

Fig. 2. Example of LSP forwarding table configurations able to avoid and produce contentions.

same input wavelength, which we refer to as grouping index δ.

Figure 3 shows the Packet Loss Rate (PLR) changing the percentage of the grouping index from $\delta = 10\%$ to $\delta = 100\%$ and the overall load from $\rho = 0.6$ to $\rho = 1$. It is clear that when $\delta = 100\%$, no contentions are possible (i.e., contention-free configuration), so the PLR is 0. When decreases, the PLR increases. The increase depends strongly on the overall load. At $\rho = 1$, the curve is practically flat, with an elbow close to $\delta = 95\%$. At $\rho = 0.7$ and $\delta = 30\%$, the PLR is less than 10^{-7}.

It must be stressed that the same behavior can be observed for any number of LSPs L contending for the same output wavelength, being the grouping index δ the key factor. We carried out several simulations varying L from 2 to 20 and the results obtained lead to the same conclusion.

2.2 The grouping wavelength assignment (GRP-WA) algorithm

The configuration of the forwarding table has a strong impact on the node performance.

In this direction, the authors in [8] propose a dynamic wavelength selection policy that is able to solve a wavelength congestion by switching the LSP to a wavelength with a higher grouping index, i.e., the wavelength at which the

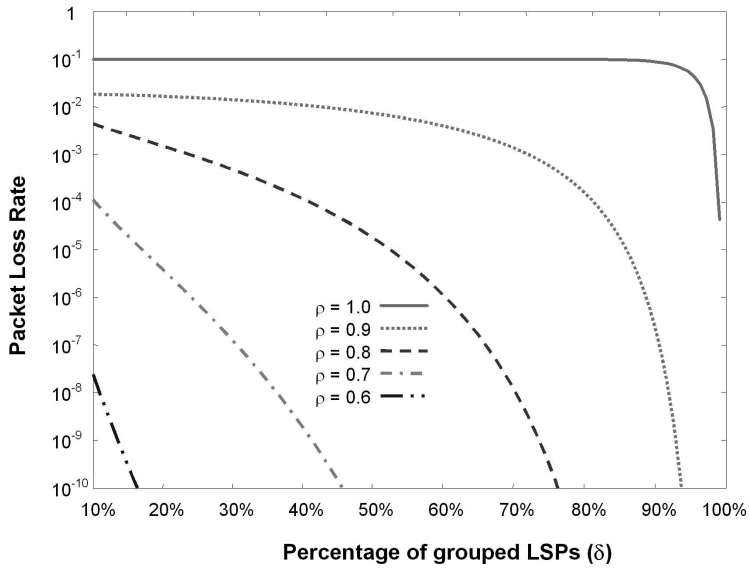

Fig. 3. Packet loss rate as a function of the relative load at different overall loads.

number of LSPs coming from the same input wavelength is highest (as close as possible to a contention-free configuration). Here we propose applying the grouping effect directly at the LSP setup. In such a way, it is possible to reduce the congestion occurrences and therefore the need to perform wavelength selection, which is the most demanding task for an optical packet switch. Previous works do not consider this issue and always assume an average situation in which the LSPs are already established and fixed in the simulations analysis [10][14].

In order to do this, we suggest that when a request to setup a new LSP arrives to a node, the procedure takes care of determining both the output port n^{out} and the output wavelength λ^{out}. While the former depends on the routing protocol, the latter may be set locally by each node using a WA algorithm which can exploit the grouping (GRP) effect and take advantage of the contention-free configuration.

The GRP-WA algorithms works as follows. At the LSP setup request, the control unit of the node recognizes the $(\lambda^{in}, n^{in}, n^{out})$-tuple of the new LSP. Therefore, it searches in the forwarding table if there is another LSP l' with the same tuple. If l' exists, it chooses the same λ^{out} assigned to l'. If not, it searches for the first wavelength λ^{out} with no assignments. If all wavelengths are already in use, the control unit balances the load of the wavelengths by assigning the least loaded one (this last step is equivalent to the BLC algorithm explained later on). For the first step, the algorithm only requires a cubic matrix of dimension $\lambda^{in} \times n^{in} \times n^{out}$, in which each entry indicates the assigned λ^{out}; a value in an entry means the presence of the l'. For the other steps, the algorithm needs to search among W values. Figure 4 describes the GRP-WA

When a request to setup an LSP l arrives

(1) Search l' with $(\lambda^{in}(l'), n^{in}(l'), n^{out}(l')) = (\lambda^{in}(l), n^{in}(l), n^{out}(l))$;

(2) If l' exists, $\lambda^{out}(l) = \lambda^{out}(l')$;

(3) If not, search for the set of wavelengths $\Lambda \in n^{out}(l)$ not used;

(4) If $\Lambda = \emptyset$, assign the least loaded wavelength.

Fig. 4. Procedure for GRP-WA setup algorithm.

procedure.

The problem here is that when a wavelength assigned to an LSP is busy (i.e., in a congestion situation), the contention resolution algorithm moves the LSP to another wavelength, which eventually breaks a contention-free configuration in the downstream node (i.e., the input wavelength of the downstream node is now different with respect to the setup configuration). Therefore, each change may require a signaling between the nodes so that the downstream node can adapt its forwarding table to the new situation. This can imply either latency or scalability concerns. To avoid this problem, we consider that any change in the forwarding table is only temporary and the LSP is switched back to the wavelength assigned at the setup as soon as the congestion disappears. Consequently the downstream node may temporarily decrease its performance during a congestion situation in the upstream node but signaling is not needed and latency is not added.

2.3 Performance evaluation

We carried out several simulations in order to evaluate the performance of the GRP-WA algorithm. To compare its results, we set up three other WA algorithms, namely **Random**, **Round-Robin**, and **Balance**:

- **Random** (RND-WA). When a request to setup a new LSP arrives, the control unit assigns a random wavelength λ^{out} of the output port n^{out}. This algorithm has very low complexity.
- **Round-Robin** (RR-WA). In this case, the control unit maintains a set of pointers per output port, each one pointing to the last assigned wavelength. At LSP setup request, the control unit increases the corresponding pointer by 1 and assigns the pointed wavelength λ^{out} of the output port n^{out}. This algorithm only requires an update of the pointers at each setup. It is similar to the First-Fit scheme extensive adopted in many WA solutions (for instance [13]).
- **Balance** (BLC-WA). In this case, we assume that the setup request indi-

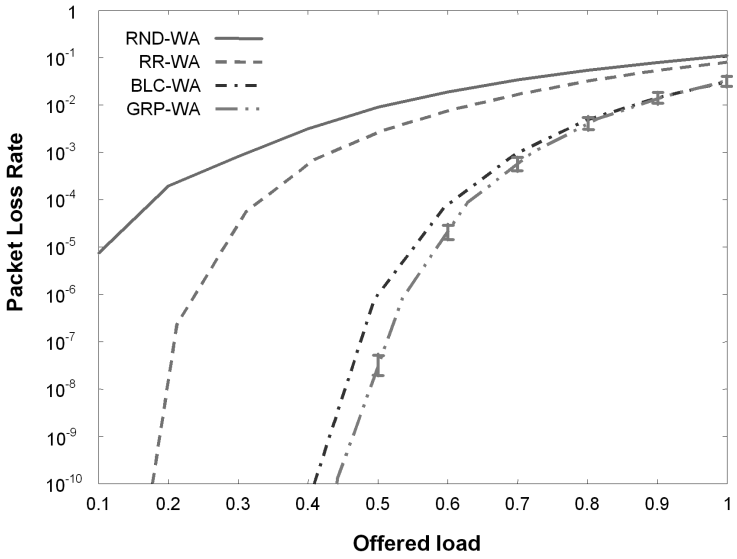

Fig. 5. Packet loss rate as a function of the overall load comparing the RND-WA, RR-WA, BLC-WA, and GRP-WA algorithms.

cates the average load of the LSP. The control unit uses this information to maintain a matrix $\mathbf{V}$, where each entry $\mathbf{V}_{i,j}$ indicates the overall load of the output wavelength i of output port j. At LSP setup request, the control unit assigns the wavelength λ^{out} of the output port n^{out} with the minimum overall load. If W is the number of wavelengths on a given output fiber, a search of the minimum among W values is required.

The simulator (described in Appendix A) is set up with $N = 4$, $W = 16$, $C = 10$ Gbps, $L = 10$, and a degenerate buffer $\mathbf{Q}_{16}$.

Two scenarios are considered.

In the first scenario we adopt a static wavelength policy and evaluate the PLR. The granularity of the buffer is $D = 0.4$, which is the optimal value for static approach [9]. Figure 5 shows the PLR as a function of the overall load, comparing the RND-WA, RR-WA, BLC-WA, and GRP-WA algorithms. As an example of result correctness, we show confidence intervals for the GRP-WA curve but not for the others for readability reasons. We can see that the BLC-WA and GRP-WA algorithms outperform the other strategies. Confirming our observations, the performance improves even more when the GRP-WA algorithm is used under light and moderate traffic load. Under higher loads, GRP-WA behaves like BLC-WA since there is less probability of grouping the LSPs or of finding free wavelengths.

Further simulations, not presented here due to lack of space, showed that whether L is changed from 2 to 20 or the traffic profile is changed, GRP-WA always shows the best PLR.

Table 1
PLR, FO and OS comparing the WA procedures with a load of 0.8

Procedure	**PLR**	**FO**	**OS**
RND-WA	$6.12\ 10^{-4}$	30.41%	3.84%
RR-WA	$6.04\ 10^{-4}$	29.28%	3.75%
BLC-WA	$5.98\ 10^{-4}$	21.85%	3.01%
GRP-WA	$5.91\ 10^{-4}$	14.97%	2.05%

In the second scenario the objective is to compare the WA procedures under a dynamic wavelength policy. We use the simple RRWS algorithm proposed in [8] which, when a wavelength is congested, switches the LSP to the first available wavelength searching in a round-robin fashion. The granularity of the buffer is $D = 0.2$, which is the optimal value for the RRWS algorithm [8]. For this scenario we evaluate the PLR as well as the Forwarding Opacity (FO) and the Out-of-Sequence packet (OS) measures. FO gives an estimation of the overload complexity measuring the percentage of time in which the RRWS algorithm performs the wavelength search step; OS gives an estimation of the packet delay measuring the percentage of out-of-sequence packets at the output of the node. The details of these measures are presented in Appendix A.

Table 1 shows the PLR, FO and OS measures with an overall load of 0.8 comparing the RND-WA, RR-WA, BLC-WA, and GRP-WA algorithms. As we can see in the table, the PLRs of the WA procedures do not show significant differences. The important results here are FO and OS. The obtained results indicate that both measures are affected from a non-optimal LSP setup. Indeed RND-WA, RR-WA and BLC-WA require more updating of the forwarding tables than GRP-WA. This increases the control complexity as well as the probability of breaking the correct packet sequence. In fact, GRP-WA obtains the lowest OS.

3 QoS provisioning

The technology limitation of the optical buffer has led to significant research efforts in recent years dealing with the design of simple contention resolution policies able to provide QoS differentiation. The impossibility of pre-emptying packets that are already buffered makes it unfeasible to implement conventional fair queuing scheduling commonly used in electrical switches. Furthermore, QoS schemes must be kept very simple to be effective in OPS where each node must be able to schedule tens of Tbit/s.

The mechanisms proposed in the literature for providing QoS in OPS net-

works use some form of resource reservation (either a buffer threshold [14] or wavelength threshold [15]), offset time differentiation [16] or hybrid electrical/optical buffers [17]. The first method does not offer good enough results –for instance in [14] the PLR for low priority class is 10^{-2} with a load of 0.8. To achieve acceptable levels of PLR with this method, the scheduling requires very high computational complexity or very large optical memories. The offset time differentiation method shows good results when applied to optical burst switching [11], in which bursts comprise several packets. However, it does not seem effective in OPS, in which the overhead of the control packets could introduce considerable bandwidth wastage. Finally, the hybrid E/O buffer method is not scalable with the bit-rate since electronic devices cannot keep up with the speed of optical links and the E/O bottleneck is maintained.

In this paper, we propose a novel strategy that is able to improve the switch performance and provide the required QoS. It consists in defining different OPS service categories (as was done in ATM networks) and the strategy is based on the fact that in a QoS environment it is not practical to provide the best handling to a service that does not really require it. Therefore, if a set of K categories of service is available in the network, each category should be handled according to its requirements. For this reason we suggest implementing a set of K different handlings (i.e., algorithms) in the node. When a packet belonging to an LSP with category i arrives at the node, the control unit will execute the corresponding algorithm i to forward the packet which guarantees only the required services. We refer to this technique as the *Service Category-to-Wavelength Selection* (SCWS) technique.

3.1 Service Category-to-Wavelength Selection technique

To study and evaluate our proposal, we defined a case study based on the following three categories of service:

- **Best Effort** (BE) with no specific QoS requirements.
- **Loss Sensitive** (LS) for multimedia broadcasting applications, which requires bounded losses.
- **Real Time** (RT) for interactive applications, which requires strict performance (very low PLR and very short delay).

We hence define three different wavelength selection algorithms, which will be implemented in the control unit. These algorithms are the following:

- **Two-State Wavelength Selection** (TSWS) for supporting the BE category of service.
- **Loss Bounding Wavelength Selection** (LBWS) for supporting the LS category of service. It can also support the BE category when there are low

When a packet belonging to LSP l arrives:

(1) Look-up the forwarding table to determine $n^{out}(l)$ and wavelengths $\lambda_0^{out}(l)$ and $\lambda_1^{out}(l)$ assigned to l;

(2) Determine the minimum delay D_j available for $\lambda_0^{out}(l)$ and $\lambda_1^{out}(l)$;

(a) If $D_j > D_M$ (i.e., both wavelength are busy), the packet is lost;

(b) If not, select the FDL j to send the packet to.

Fig. 6. TSWS algorithms.

LS connection demands.

- **Sequence Keeping Wavelength Selection** (SKWS) for supporting the RT category of service. It can also support the BE and LS categories when there are low RT connection demands.

The aim of the TSWS algorithm is to reduce the control overload (low FO) whilst maintaining an acceptable level of the PLR for the BE traffic. This algorithm tries to improve the performance of the static approach by assigning two wavelengths to the LSP during the setup procedure. The GRP-WA algorithm is hence executed twice: the second time, the first assigned wavelength is excluded from the search in order to avoid choosing the same wavelength twice. The wavelength assignments are kept constant throughout the LSP life and single packets are always forwarded to the least congested wavelength. This means that the wavelength searching step of the wavelength selection algorithm is never needed (FO is always 0%). Consequently, the complexity of this algorithm mainly concerns the setup procedure, while a simple comparison between two values is required for each incoming packet. Figure 6 shows the steps to follow when running this algorithm.

The aim of the LBWS algorithm is to achieve a bounded PLR. The initial wavelength assignment may change if the LSP experiences a PLR above a predetermined value R (*required PLR*). For this purpose, a window T is defined. Every T the algorithm computes the PLR of each LSP. In order to be viable, the PLRs can be simply estimated using a sampling technique. Once PLRs are obtained, they are then sorted in descending order. Starting from the higher value, the algorithm compares the PLRs with R; if it is higher, a new GRP-WA algorithm is executed to reassign the LSP to another wavelength. The complexity of LBWS depends on the number of LSPs L and on the duration of the window T. Indeed, ordering $L \times W \times N$ values is required every T. Hence, T is a parameter to be set carefully and can affect the node performance: high values may not guarantee the required PLR; on the other hand, low values can increase the control overload with an extreme situation of executing a new GRP-WA algorithm for each incoming packet. Finally, it

When a packet belonging to LSP l arrives:

(1) Look-up the forwarding table to determine $n^{out}(l)$ and $\lambda^{out}(l)$ assigned to l;

(2) Determine the minimum delay D_j available for $\lambda^{out}(l)$;

(a) If $D_j > D_M$, the packet is lost;
(b) If $D_j \leq D_M$, select the FDL j to send the packet to.

Every window T:

(1) Determine the PLR of the LSPs: PLR^T;

(2) Order (descending) the PLR^T;

(3) From $l = 1$ to $L \times W \times N$:

(a) If $PLR_l^T > R$, execute a new GRP algorithm

Fig. 7. LBWS algorithms.

is important to note that the value of R can be different from one LSP to another since their requirements may be different. For the sake of simplicity, in this work we assume the same value for all LSPs that use this algorithm. Figure 7 shows the steps to follow when running this algorithm.

The SKWS algorithm was originally proposed in [10]. Its aim is to achieve the required level of PLR by maximizing the resource utilization and throughput. At the same time, SKWS needs to control the delay preserving the correct packet sequence belonging to the same LSP. Indeed, since very short optical buffers are available in OPS networks, the delay is only due to the propagation delay and to rebuild the original information at the edge of the optical network. The latter may introduce considerable delays if extensive reordering operations are needed due to the out-of-order delivery of the packets [18].

For the purpose of this work, given a stream of ordered packets at the switch input, we define the packet i to be out-of-order when the first bit of packet i leaves the node before the last bit of packet $i-1$. Less restrictive cases are more difficult to control, especially when a cascade of nodes is considered. In fact, taking into account that in general the optical packets can aggregate more than one IP packet, the relative position of subsequent IP packets included in two subsequent optical packets cannot be controlled if overlapping is permitted. Therefore, a strict sequence keeping (i.e., avoiding packet overlapping) represents the only procedure that ensures the maintenance of sequence both at the optical packet level and at the IP packet level. Consequently, in this work we have adopted this restrictive case.

When a packet with duration d belonging to LSP l arrives at time t:

(1) Look-up the forwarding table to determine $n^{out}(l)$, $\lambda^{out}(l)$ and $t^{out}(l)$ assigned to l;

(2) If $t^{out}(l) \geq t$, the previous packet has already left the node:
   (a) If $\lambda^{out}(l)$ is busy, search for the set of wavelength $\Lambda \in n^{out}(l)$ not busy:
   (b) If $\Lambda = \emptyset$, the packet is lost;
   (c) If $\Lambda \neq \emptyset$, determine the wavelength λ with the minimum delay D_j and select the FDL j to send the packet on;

(3) If $t^{out}(l) < t$, the previous packet is still in the node:
   (a) Calculate the minimum delay D_{min} to add to the packet to preserve the order $D_{min} = \left\lceil \frac{t^{out}(l)-t}{D} \right\rceil D$;
   (b) If $\lambda^{out}(l)$ cannot provide D_{min}, search for the set of FDL F able to provide a delay $D_j \geq D_{min}$
   (c) If $F = \emptyset$, the packet is lost;
   (d) If $F \neq \emptyset$, select the FDL with minimum D_j, if more than one is available, select the wavelength λ that introduces the minimum gap between two subsequent packets;

(4) Update the forwarding table $t^{out}(l) = t + d + D_j$, $\lambda^{out} = \lambda$.

Fig. 8. SKWS algorithms.

To keep the correct packet order, the control unit stores the time-stamps t^{out} (one per each LSP) at which the last bit of the last packet is scheduled to leave the node. This time is calculated as the sum of the packet arrival time, its duration and the delay assigned in the buffer. When a packet belonging to the LSP l arrives, the control unit recalls the time $t^{out}(l)$ and determines whether the new packet needs an additional delay to keep the order. This delay must be at least as long as the residual transmission time of the previous packet belonging to the same LSP l. Due to the discrete number of delays provided by the optical buffer, the additional delay is calculated as the integer multiple of D greater than $t^{out}(l)$.

Figure 8 shows the steps to follow when running the SKWS algorithm. Regarding the complexity, at maximum both point 2 and point 3 require a search for a minimum among the W values. The difference is that point 3 requires the calculation of the additional delay D_{min} and the minimum is relative to this value.

In this section, the algorithms are evaluated separately in order to find their specific characteristics. Afterwards, we integrate them in the same switch and evaluate the SCWS technique.

In the following figures, we set up the simulator (described in Appendix A) with $N = 4$, $W = 16$, $C = 10$ Gbps, and $L = 3$. The buffer configuration is a degenerate buffer $\mathbf{Q}_8$ (i.e., the length is $B = 8$) except for SKWS, which uses a shorter buffer $\mathbf{Q}_6$. The offered load is $\rho = 0.8$ except for LBWS, where it is $\rho = 0.6$ because it is not possible to bound the PLR of a high amount of traffic whilst maintaining an acceptable control complexity. R is set to 10^{-4} and T to 20 D, which are reasonable values offering a good trade-off between complexity and PLR.

Figure 9 shows the simulation results when the three algorithms are evaluated independently. Figure 9a plots the PLR as a function of D normalized to the average packet duration, comparing the TSWS, LBWS and SKWS algorithms. In this figure we can see that SKWS achieves the best PLR of 10^{-6} with $D = 1.2$. Contrarily to the usual concave behavior shown by other algorithms, LBWS exhibits constant values of less than 10^{-4}, which is the value set as required. TSWS shows the worst PLR but it is important to note that its aim is to have low control complexity. Figure 9b plots the Forwarding Opacity (FO) measure, comparing TSWS, LBWS, and SKWS. It is clear that SKWS imposes the highest overload on the switch control, while LBWS shows low computational requirements with values close to 4%. The LBWS curve indicates that keeping PLR bounded requires less computations for values of D ranging between $D = 1$ and $D = 1.4$, with a minimum at $D = 1.2$. Finally, TSWA does not need to reconfigure its LSP-to-wavelength mapping, so FO is always 0%. Figure 9c shows the percentage of Out-of-Sequence packets (OS) comparing the TSWS, LBWS, and SKWS algorithms. As expected, SKWS maintains the correct sequence delivery. LBWS shows values of around $2 \div 2.5\%$, while TSWS exhibits a concave behavior with a maximum of 3.7% at $D = 0.7$.

The results obtained previously assess the goodness of the proposed algorithms, indicating that their aims have been fully accomplished: TSWS imposes low control overload and reaches acceptable PLR; LBWS requires low control overload and is able to guarantee a bounded PLR; finally, SKWS requires a high control overload but achieves very good PLR, maintaining the correct order of the packet sequence. The next step is hence the integration of these algorithms in the same control unit and the verification of the mutual impacts on the performance measures.

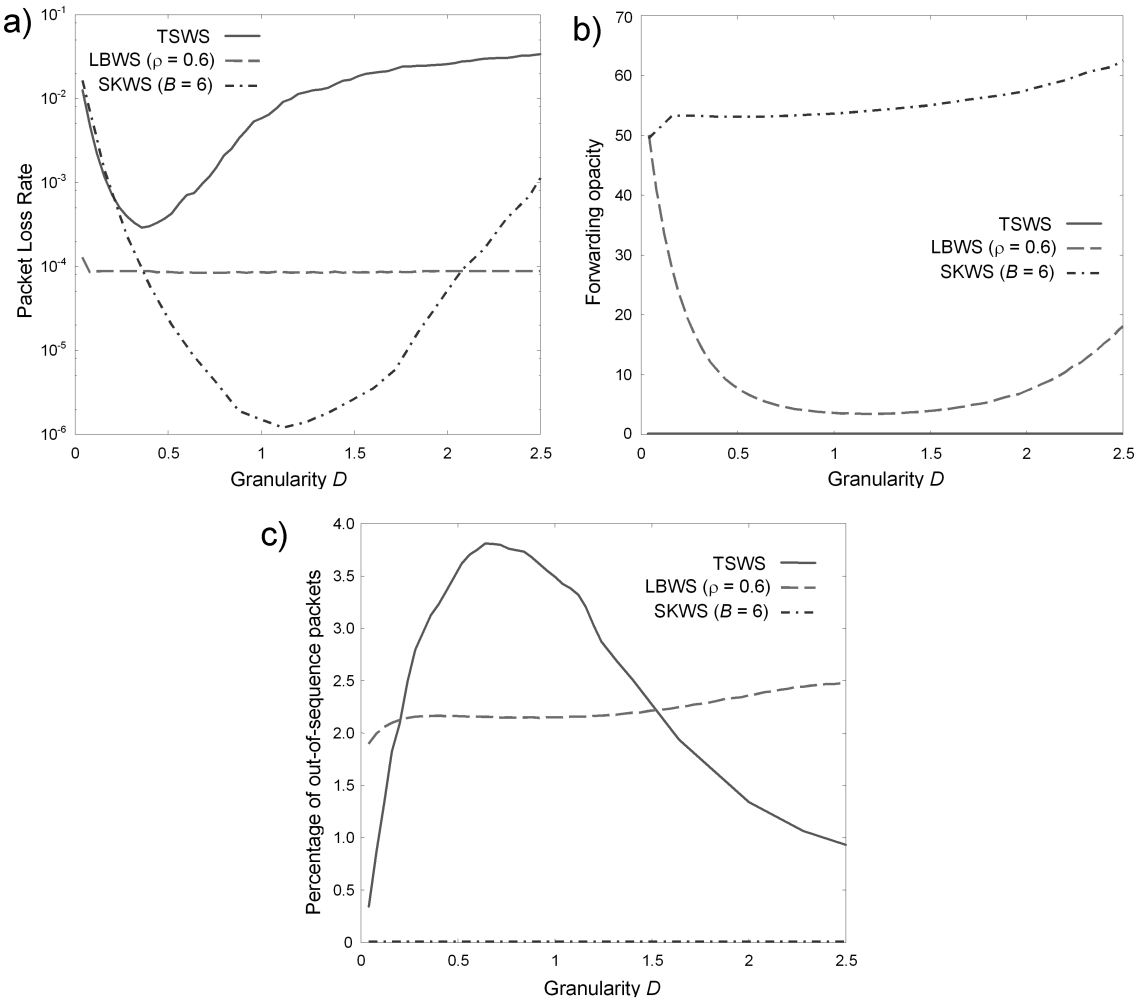

Fig. 9. a) Packet loss rate, b) Forwarding opacity, and c) Out-of-sequence packets as a function of D normalized to the average packet duration, comparing TSWS, LBWS and SKWS.

The integration is not trivial because the performances of the algorithms are not aligned, i.e., they show the best results with different values of the fiber granularity D: the optimum D for LBWS and SKWS is 1.2 while for TSWS it is 0.4 (see Figure 9a and Figure 9b). Thus, we design an ad-hoc optical buffer architecture to achieve the perfect performance alignment between the algorithms.

3.3 Ad-hoc optical buffer architecture

The idea comes from the observation that the ratio of these optimum values of D (1.2 and 0.4) is exactly 3. Exhaustive simulations (not presented here due to lack of space) show that this peculiar factor of 3 is valid for any traffic profile. It only depends on the average packet size and on the transmission

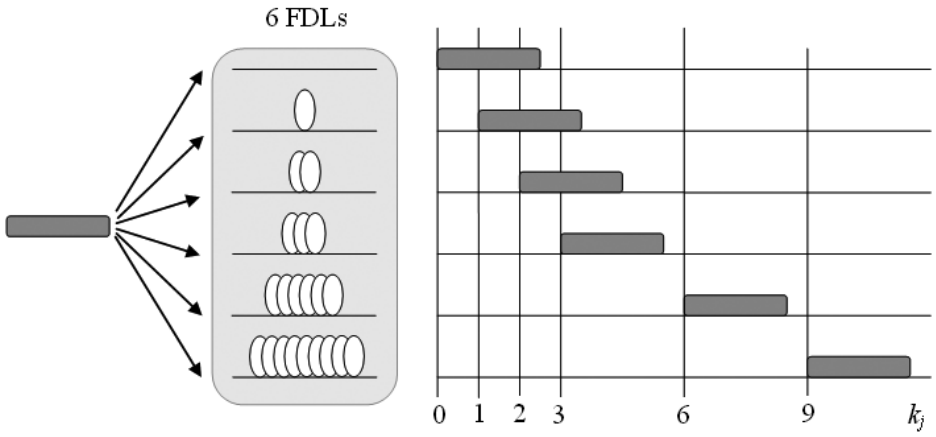

Fig. 10. Non-degenerate buffer configuration with 6 FDLs. BE packets can use delays $\{0, D, 2D, 3D\}$, while the RT and LS packets can use delays $\{0, 3D, 6D, 9D\}$

bit-rate.

Based on this factor, the integration of the different wavelength selection algorithms can be done using the following buffer architecture. Firstly, we fix $D = 0.4$ and set up two degenerate buffers: $\mathbf{Q}'$ with $D_j = jD$ delays and length B' for the BE packets, and $\mathbf{Q}''$ with $D_j = 3jD$ delays and length B'' for RT and LS packets. Thus, these buffers are merged in a non-degenerate buffer $\mathbf{Q} = \mathbf{Q}' \cup \mathbf{Q}''$ in such a way that the delays that are common in $\mathbf{Q}'$ and $\mathbf{Q}''$ are available for any category. Figure 10 shows an example with $B' = B''$ $= 4$, and a resulting length $B = 6$ of buffer $\mathbf{Q}$.

It is important to note that in our simulations we use a feed-forward buffer configuration, but the concept of merging several degenerate buffers in a single non-degenerate structure is also valid for a feedback configuration. It is clear that different values of fiber granularity may be required and should be determined for each algorithm.

For the evaluation under multi-category, we set $N = 4$, $W = 16$, $B = 6$ $C = 10$ Gbps, $\rho = 0.8$, $L = 3$, and finally the required PLR and measure window for LS packets to $R = 10^{-5}$ and $T = 20$ D, respectively. Regarding the distribution of traffic, in Figure 11, Table 2 and Figure 12 we assume that 50% of the LSPs transport BE packets, 30% transport RT packets, and the rest LS packets. In Figure 13 we analyze the PLR changing this distribution.

Figure 11 plots the PLR for the entire system as a function of D normalized to the average packet duration. In the figure, we include the secondary x-axis which indicates the granularity perceived by SKWS and LBWS algorithms (exactly 3 times D). As expected, the performance of the three categories of service became aligned, and the optimum value was obtained for $D = 0.4$. Hence, we use this value to obtain the following results.

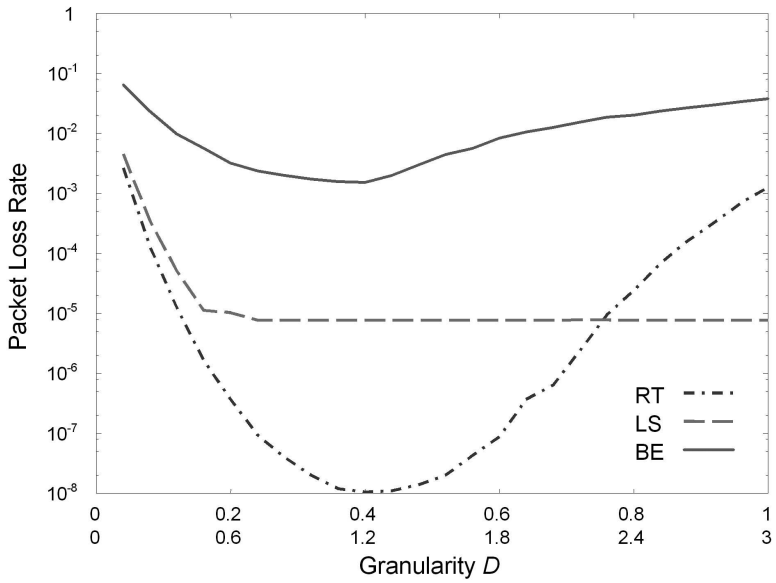

Fig. 11. Packet loss rate as a function of D normalized to the average packet duration.

In Table 2, we compare the SCWS technique with the *Empty Queue Wavelength Selection* (EQWS) algorithm [9] –the best performed dynamic wavelength selection algorithm– and the *Minimum Gap* (MINGAP) algorithm [12] –the best performed connectionless algorithm. Both EQWS and MINGAP use the buffer threshold approach [14] to provide QoS (the values of D and of thresholds are those providing the lowest PLRs).

The results show that the SCWS technique provides the lowest PLR for both LS and BE traffic. Moreover, as expected, the higher control complexity is required to forward the RT traffic (FO is 66.14%), while LS and BE impose low overload (5.93 and 0% respectively). In contrast, MINGAP imposes the same (very high) FO for any category, while EQWS requires higher FO for BE traffic, which is obvious nonsense. Furthermore, the packet sequence of RT traffic is preserved using the SCWS technique, while it reaches 2 and 5% using EQWS and MINGAP respectively. Previous studies [19] [20] confirm that even a small percentage of out-of-sequence (like that caused by the EQWS algorithm) may have a harmful impact on the network performance. We must also consider that this percentage is counted at the output of a single node; by assuming n nodes in series along a path this percentage increases accordingly.

Figure 12 plots the PLR as a function of the buffer depth B and of the number of wavelengths W for any category. The results indicate that a significant improvement in the performance can be obtained with a small increase in the number of FDLs B of buffer $\mathbf{Q}$ (Figure 12a) as well as in the number of wavelengths (Figure 12b).

Finally, Figure 13 shows the PLR, changing the percentage of the relative

Table 2
PLR, FO and OS comparing the SCWS technique with EQWS and MINGAP, both adopting a buffer threshold technique.

	SCWS		
Category	**PLR**	**FO**	**OS**
RT	$1.08\ 10^{-8}$	66.14%	0%
LS	$7.68\ 10^{-6}$	5.93%	1.76%
BE	$1.55\ 10^{-3}$	0%	3.29%
	EQWS		
RT	$3.00\ 10^{-8}$	16.20%	2.02%
LS	$2.75\ 10^{-4}$	30.82%	2.33%
BE	$5.24\ 10^{-2}$	52.51%	3.41%
	MINGAP		
RT	$< 10^{-9}$	81.33%	5.39%
LS	$9.78\ 10^{-4}$	81.05%	5.03%
BE	$3.96\ 10^{-3}$	80.92%	4.62%

load between the RT and BE traffic while maintaining the relative load of LS traffic fixed to 20%. This means that, for instance, when the percentage of RT is 20%, the percentage of BE is 60%. We can see that the PLR of LS cannot be guaranteed if there is a high percentage of RT traffic (i.e., more than 60%). On the other hand, if RT is not present, BE traffic is not able to fully exploit the node capacity and the PLR remains relatively high. A way to improve the performance of the BE traffic when RT and LS have low loads is to apply the SKWS algorithm also to some BE LSPs. In this case, a smart algorithm must be developed in order to decide when, which, and how many BE LSPs can be forwarded according to the SKWS algorithm. This study is not developed here and is left for future research.

4 Summary and conclusion

In this paper, we considered a connection-oriented OPS node and addressed two different problems: the problem of establishing the optical LSP and properly configuring the forwarding table at the node by means of a smart wavelength assignment procedure and the problem of providing QoS.

For the first problem, a procedure called GRP-WA was proposed. The results demonstrated that considerable switch performance improvements can be obtained by grouping the contention-free flows (i.e., flows coming from the same input wavelength). For example, for the case of a lightly and moderate loaded

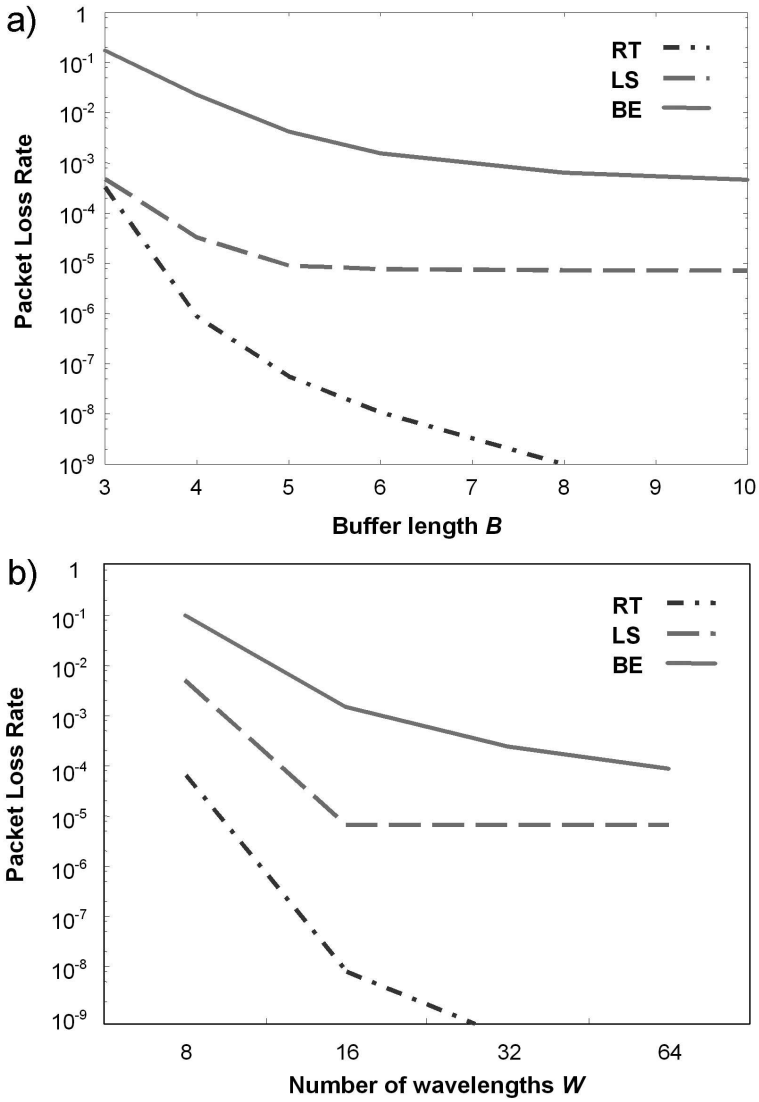

Fig. 12. Packet loss rate as a function of a) the buffer length B, b) the number of wavelengths W.

node and under static wavelength policy, the GRP-WA algorithm yields a PLR of one order of magnitude lower than balancing the LSP load. Under a dynamic wavelength policy the benefits mainly yields in a complexity reduction: in fact, GRP-WA requires less forwarding table updating (lower FO) and fewer packets are delivered out-of-sequence (lower OS).

For the second problem, the novel SCWS technique was proposed to provide QoS. We carried out a case study with three different OPS service categories and designed three different wavelength selection algorithms. The key point of the study was the design of an ad-hoc buffer structure able to coordinate the behavior of the different algorithms and optimize the node performance. The results obtained highlight its effectiveness as well as its improvements compared with previous works.

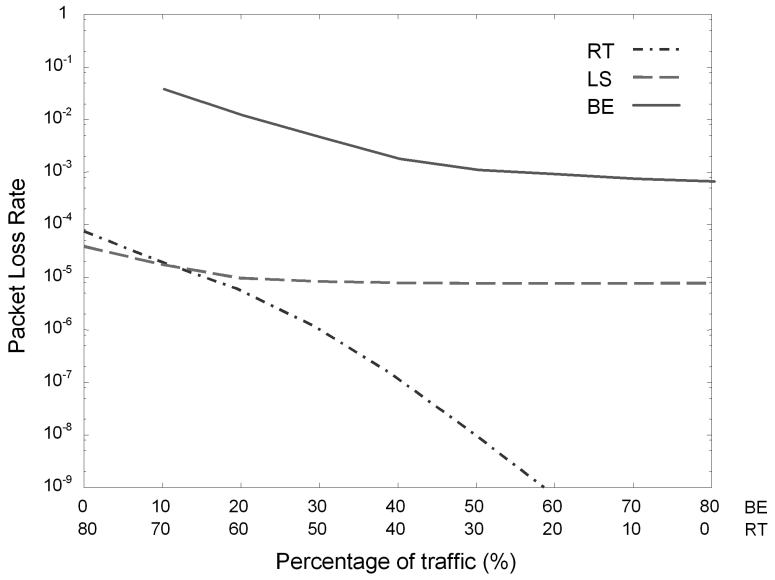

Fig. 13. Packet loss rate as a function of traffic relative load percentage.

Several further improvements are currently underway. We are working on the application of the SCWS technique to a feedback buffer configuration. As we did for the feed-forward buffer configuration addressed in this paper, the objective will be to find the optimal fiber granularity for the different wavelength selection algorithms and therefore merge the degenerate buffers into a single non-degenerate structure.

Further studies also focus on designing a service differentiation mechanism in the GRP-WA procedure, which, integrated with the SCWS technique, may obtain a more flexible environment. For example, we think in defining two or more LSP classes and set them up with different priority preferences and/or blocking probabilities.

Finally, in this paper we have only presented the case of a single node scenario, but the SCWS technique opens up interesting developments in the routing problem for a whole network scenario.

A Simulation environment

In this appendix we present the simulation environment. It consists of an event-driven program which simulates the behavior of a single optical packet switch. We do not deal with implementation issues but with performance analysis, so we consider that a non-blocking OPS switch architecture with full wavelength conversion capabilities is available. The optical switch parameters are:

- N, the number of input and output fibers;

- W, the number of wavelengths per fiber;
- C, the transmission bit-rate;
- D, the delay granularity of the FDLs;
- $\mathbf{Q}_B$, the set of possible delays of B FDLs where D_M is the maximum available delay; if the delays are consecutive, the buffer is said to be *degenerate* (in this case $M = B - 1$), otherwise, it is *non-degenerate* [5] (in this case the value of M depends on the buffer configuration);
- L, the average number of LSPs per input wavelength.
- ρ, the offered load which is the same for any input and output wavelength (i.e., uniform distribution).

The distribution of the LSPs follows an exponential model: both the interarrival time and the holding time are exponential distributed. The mean value of the interarrival times, connection duration, and required bandwidth are selected accordingly to generate the required offered load ρ. Using the exponential assumptions is a consolidated approach for the LSP distribution (see for instance [13]) because neither traffic measurements nor statistical analysis are currently available.

For the packet model, we consider recent studies on traffic measurement (for instance [21]). They illustrate that aggregated traffic in Internet during a long period of observation is not self-similar as discovered ten years ago but is more close to the Poisson model. Since the traffic in the OPS network is expected to be even more aggregated than that observed in the Internet, we assume that the interarrival time is exponential distributed with an average that depends on the LSP bandwidth. The packets have an exponential distributed size with average and minimum lengths of 500 and 40 bytes respectively. The number of simulated packets is chosen large enough to reach steady-state results and a 95% confidence interval is calculated.

We define the following measures to evaluate the node performance:

- *Average Packet Loss Rate* (PLR). This is the usual performance measure for packet switches and also indicates the capability of an algorithm to reduce the congestion situation.
- *Out-of-sequence packets* (OS). This measure indicates the percentage of out-of-sequence packets belonging to the same LSP. The lower the percentage, the lower the amount of packets to be reordered at the destination.
- *Forwarding opacity* (FO). This is measured as the percentage of packets that are forwarded searching for a new wavelength over the total number of simulated packets. The resulting value estimates the overload on the switch control function. The higher the percentage, the higher the overload.

The aim is therefore to lower all three metrics consistently with the QoS traffic requirements.

Acknowledgment

This work was partly funded by the European Union under the IP NOBEL project (IST FP6-506760) and partly by the MCYT (Spanish Ministry of Science and Technology) under the contract TEC2005-08051-C03-01/TCM (CATARO project).

References

[1] W.D. Sincoskie, "Broadband packet switching: a personal perspective", *IEEE Commun. Mag.*, vol. 40, no. 7, Jul. 2002, pp. 54–66.

[2] S. Spadaro, J. Solé-Pareta, D. Careglio, K. Wajda, A. Szymanski, "Positioning of the RPR standard in contemporary operator environments", *IEEE Network*, vol. 18, no. 2, Mar. 2004, pp. 35–40.

[3] L. Dittman et al., "The IST project DAVID: a viable approach towards optical packet switching", *IEEE J. Select. Areas Commun.*, vol. 21, no. 9, Sep. 2003, pp. 1026–1040.

[4] T.S. El-Bawab, J.-D. Shin, "Optical packet switching in core networks: between vision and reality", *IEEE Commun. Mag.*, vol. 40, no. 9, Sep. 2002, pp. 60–65.

[5] D.K. Hunter, M.C. Chia, I. Andonovic, "Buffering in optical packet switches", *IEEE/OSA J. Lightwave Technol.*, vol. 16, no. 12, Dec. 1998, pp. 2081-2094.

[6] D. Chiaroni et al., "First demonstration of an asynchronous optical packet switching matrix prototype for multiterabit-class routers/switches", in *Proc. 27th Eur. Conf. Optical Commun.* (ECOC 2001), Amsterdam, The Netherlands, Oct. 2001.

[7] L. Tančevski et al., "Optical routing of asynchronous, variable length packets", *IEEE J. Select. Areas Commun.*, vol. 18, no. 10, Oct. 2000, pp. 2084-2093.

[8] F. Callegati, W. Cerroni, C. Raffaelli, P. Zaffoni, "MLPS over optical packet switching", in *Proc. Thyrrenian International Workshop on Digital Communications* (IWDC2001), Taormina, Italy, Sep. 2001.

[9] F. Callegati, W. Cerroni, C. Raffaelli, P. Zaffoni, "Dynamic wavelength assignment in MPLS optical packet switches", *Optical Network Mag.*, vol. 4, no. 5, Sep. 2003, pp. 41-51.

[10] F. Callegati, D. Careglio, W. Cerroni, J. Solé-Pareta, C. Raffaelli, P. Zaffoni, "Keeping the packet sequence in optical packet-switched networks", in *Optical Switching and Networking Journal*, vol. 2, no. 3, pp. 137-147, Nov. 2005.

[11] M. Yoo, C. Qiao, S. Dixit, "QoS performance of optical burst switching in IP-over-WDM networks", *IEEE J. Select. Areas Commun.*, vol. 18, no. 10, Oct. 2000, pp. 2062-2071.

[12] F. Callegati, W. Cerroni, G. Corazza, "Optimization of wavelength allocation in WDM optical buffers", *Optical Network Mag.*, vol. 2, no. 6, Nov. 2001, pp. 66-72.

[13] X. Chu, B. Li, "Dynamic routing and wavelength assignment in presence of wavelength conversion for all-optical networks", *IEEE/ACM Transactions on Networking*, vol. 13, no. 3, Jun. 2005, pp. 704–715.

[14] F. Callegati, W. Cerroni, C. Raffaelli, P. Zaffoni, "Wavelength and time domain exploitation for QoS management in optical packet switches", *Computer Networks*, vol. 44, no. 4, Mar. 2004, pp. 569-582.

[15] F. Callegati, G. Corazza, C. Raffaelli, "Exploitation of DWDM for optical packet switching with QoS guarantees", *IEEE J. Select. Areas Commun.*, vol. 20, no. 1, Jan. 2002, pp. 190-201.

[16] C. Develder, J. Cheyns, M. Pickavet, P. Demeester, "Service differentiation mechanisms for variable length packets in an optical switch with recirculating FDL buffer", in *Proc. Photonic in Switching* (PS 2003), Versailles, France, Sep. 2003.

[17] S. Bjornstad, N. Stol, D.R. Hjelme "Quality of service in optical packet switched DWDM transport networks", in *Proc. SPIE*, vol. 4910, *Xie, S., Qiao, C., Chur Chung, Y (eds.): Optical networking II*, Sep. 2002, pp. 63-74.

[18] M. Laor, L. Gendel, "The effect of packet reordering in a backbone link on application throughput", *IEEE Network*, vol. 16, no. 5, Sep. 2002, pp. 28-36.

[19] J.C.R. Bennett, C. Patridge, "Packet reordering is not a pathological network behavior", *IEEE/ACM Trans. Networking*, vol. 7, no. 6, Dec. 1999, pp. 789-798.

[20] S. Jaiswal, G. Iannacone, C. Diot, J. Kurose, D. Towsley, "Measurement and classificatiion of out-of-sequence packets in a tier-1 IP backbone", in *Proc. IEEE Infocom 2003*, San Francisco, CA, vol. 2, Mar. 2003, pp. 1199-1209.

[21] T. Karagiannis, M. Molle, M. Faloutsos, A. Broido, "A nonstationary Poisson view of Internet traffic", in *Proc. IEEE Infocom 2004*, Hong Kong, Mar. 2004.