Tandem Architecture for Photonic Packet Switches

Maurizio Casoni and Carla Raffaelli

Abstract: **A new switch architecture is presented to enhance output queuing in photonic packet switches. Its application is for a packet switching environment based on the optical transport of fixed length packets. This architecture consists of a couple of cascading switching elements with output queuing, whose buffer capacity is limited by photonic technology. The introduction of a suitable buffer management allows a very good and balanced exploitation of the available optical memories, realized with fiber delay lines. In particular, packet loss performance is here evaluated showing the improvement with respect to the single switch and a way to design large optical switches is shown in order to meet broadband network requirements.**

Index Terms: **Broadband switching architectures, photonic switching, output queuing, fibre delay lines.**

I. INTRODUCTION

The current telecommunication environment is evolving towards increasingly heterogeneous interconnected networks with envisaged growth for existing and new advanced services, such that a very significant increase in bandwidth demand is expected for the years to come [1]. In relation to the mentioned trends, future networks and nodes are required to reach beyond the possibilities of electronic implementations. Optical technologies is an attractive solution for supporting not only transmission but also routing and multiplexing/demultiplexing of information flows. As a first step in this direction Optical Cross Connects (OXCs) and Optical Add-Drop Multiplexers (OADMs) have been studied and developed for all-optical Networking [2]. The range of future services is also very different in terms of required channel capacity, generated traffic pattern and connection duration. The introduction of the Asynchronous Transfer Mode (ATM) promises for a unified telecommunication infrastructure for flexible management of all kinds of information. Circuit switching, in fact, lacks the flexibility in terms of bandwidth allocation and granularity that is typical of packet switching and in particular of the ATM. All optical ATM switching and optical packet switching (see [2] for a survey) started to be investigated some years ago. In Europe, in the framework of the RACE II program, the project ATMOS has investigated and proven the feasibility of all optical ATM switches [3]. More recently, the KEOPS project has been developed in the framework of the European ACTS program with the aim of defining an optical network layer, based on packet transfer mode [4]. These studies have been carried on accordingly to the requirements of optical technology, capable of providing a fully transparent transport infrastructure which is thought to act as a very high capacity backbone interconnecting, for example, ATM or IP networks [5], [6], and to enhance flexibility and scalability.

Optical packet switching architectures have been proposed in these frameworks [3], relying on the so-called broadcast-and-select (BS) scheme. The BS matrix, demonstrated with the build up of a laboratory switching test-bed operating at 2.5 Gbit/s [7] and 10 Gbit/s [8], is suited for fast packet switching with the assumption of fixed duration packets, regardless of the characteristics of carried information. By properly managing a set of WDM fiber delay lines, through electronic control, output queuing can be achieved. The buffer capacity is limited by technology to a few tens of locations [9], due to the available splitting factor: this capacity is not sufficient to meet broadband switching requirements [10]. Large buffers are in fact necessary in this environment, especially in the presence of high load or concentration in multistage switch architectures used for large switch implementation [11].

A novel switch architecture is here proposed with the aim to overcome the problem of limited buffer size. The main goal of our work was to find out more effective techniques than the pure output queuing in the photonic domain in order to get larger memories. This issue is actively studied today and other proposals are coming out [12]. The solution proposed here achieves less queueing capacity than [12], but with higher granularity, such that a better trade-off between complexity and performance can be obtained.

The paper is organized as follows. Section II describes the functional behaviour of the basic switching matrix. Section III describes the proposed switch architecture based on a couple of cascading switching elements, called *tandem switch architecture*. Section IV describes in detail the control techniques adopted to improve the storage capability. Section V describes the algorithms applied to first stage queues to achieve fairness. In Section VI performance evaluations of the tandem architecture are reported and compared to the single BS switching element. Section VII shows an application example where the basic tandem switching element is used in the design of a large photonic packet switch. Finally in section VIII conclusions are drawn.

Manuscript received February 9, 1999; approved for publication by Alan Willner, Division I Editor, July 27, 1999.

M. Casoni is with the Dept. of Engineering Sciences of the University of Modena and Reggio Emilia, e-mail: mcasoni@deis.unibo.it.

C. Raffaelli is with D.E.I.S. of the University of Bologna, Viale Risorgimento 2, 40136 Bologna, Italy, e-mail: craffaelli@deis.unibo.it.

This work has been partially supported by the Commission of the European Community, ACTS Project AC043 "Keys to Optical Packet Switching" (KEOPS).

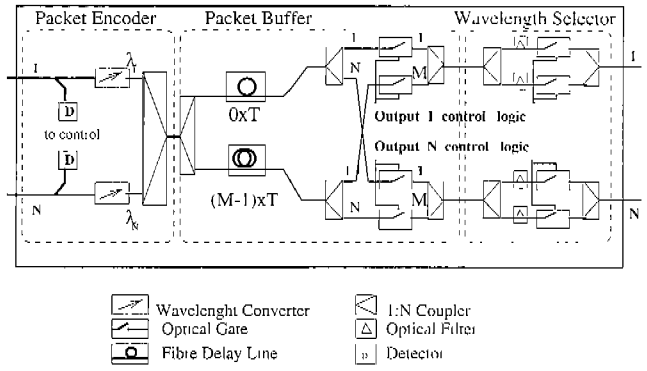

Fig. 1. The broadcast and select switching element.

II. BROADCAST & SELECT SWITCH FUNCTIONAL DESCRIPTION

The single Broadcast-and-Select Switching Element(BSSE) has been described in several previous papers both in a system [13] and in a component perspective [14], [15], [16]. It is sketched in Fig. 1 to recall its capability to realize output queuing by means of a suitable control of optical gates in the packet buffer and in the wavelength selector.

The BSSE relies on the use of wavelength encoding to achieve packet routing and on the exploitation of optical fiber delay lines (FDLs) accessed by fast optical gates to perform packet buffering and time switching. The $N \times N$ matrix, described in Fig. 1, is basically composed of three main blocks, namely the packet encoder, the packet buffer and the wavelength selector.

- The wavelength encoding block consists of a set of N all-optical wavelength converters (AOWCs), one per input and each being assigned one fixed wavelength.
- The buffering and time switching block includes M FDLs and a space switch realised using semiconductor optical amplifier gates (SOAGs).
- The wavelength selection block is based on a set of N wavelength channel selectors (WCSs) built up with optical de-multiplexers and SOAGs.

The principle of operation can be depicted as follows. Each incoming packet is assigned one wavelength by means of a wavelength converter identifying its input port, and then fed into the packet buffer. All packets are diffused to all FDLs so as to experience all possible delays achievable within the buffer. The role of the optical gates of the buffer associated to each output of the switch is to select one time slot corresponding to the appropriate delay, as determined from current traffic conditions at the input. Virtual queues with a FIFO discipline are created by using pointers associated to each outlet matrix to realise a pure output queuing architecture. All signals, at all available wavelengths are gated simultaneously. Finally the wavelength selector only discriminates one outgoing packet at a time, on the basis of its encoding wavelength, i.e., its input address. The routing and control logic drives the optical gates based on information contained in the packet header. Multi-casting, as required by future applications and services, is easily provided since the same wavelength can be selected at any time at any output port.

Four key devices and building blocks are required for the implementation of this architecture, namely the AOWC, the SOAG and the WCS.

- The AOWC can be based on a SOA operated under cross-saturation regime [7]. This device has proven to be easy to use thanks to its polarisation and wavelength insensitivity and could moreover accommodate very high bit rates. More advanced devices relying on cross-phase modulation within SOAs integrated in interferometric structure will ultimately provide some signal regeneration (extinction ratio enhancement) [17].
- The optical buffer consists of a set of M optical fibres used as delay lines, with increasing length corresponding to relative propagation times of $0 \times T$ up to $(M-1) \times T$, where T is the time slot duration [9] equal to the packet time.
- Both the buffer and the SOAGs are to be operated in a multi-wavelength regime. To avoid inter-channel cross-modulation which might occur in conventional SOAs, clamped-gain SOAs [15] are required, which maintain the gain at a constant value. Dynamic ranges of input power well above 12 dB have been recorded at 2.5 Gbit/s on first samples tested in laboratory, allowing a capacity of 16 channels in such an architecture.

The WCS could be realised by combining one optical demultiplexer to separate the wavelengths, a set of SOAGs to achieve the packet selection and one multiplexer to recombine the selected packets towards one common output port. This device takes advantage from the fast switching capability of the SOAGs.

A packet entering the switching element (SE) can be available for transmission on the addressed output with different delays, between 0 and $(M-1)T$. When multiple arrivals for the same output in the same time-slot occur, they can be scheduled for that output with different delays. The output control logic must know the delay gate and the wavelength gate to close for the right transfer, packet time per packet time. The packet buffer management can be explained underlining the relationship between the logical FIFO output buffer and the physical delay line. On a packet arrival, a counter initialized at one is associated to it and it is increased by one unit for each packet time the packet remains in the SE. Such counter indicates for each packet the number of the delay line on which the packet is available with the minimum delay and it will be used by the control logic to perform the FIFO strategy. Then the position occupied by a packet in the logical FIFO buffer is generally different from the number of the delay line where the packet is physically available. Since the number of delay lines is finite, packet loss may occur because the largest delay a packet can have is $(M-1)T$, after that it is lost. This may occur as a consequence of multiple arrivals. If the control unit works serving the delay lines according to the counters, the SE achieves a pure output queueing [10].

III. TANDEM SWITCH ARCHITECTURE

The architecture here proposed is called Tandem Switch Architecture and consists of a couple of $N \times N$ BSSEs as shown in Fig. 2. Output queuing capacity enhancement is here addressed to improve the performance with respect to the single BSSE and in particular to meet the broadband requirements in terms of packet loss probability. The possibility of cascading

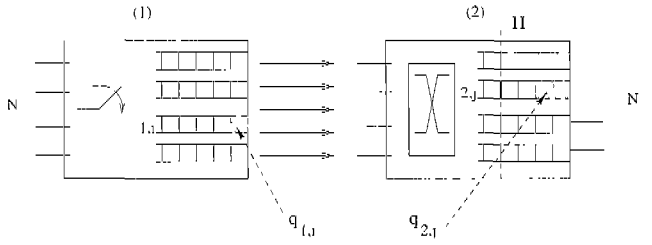

Fig. 2. The Tandem Switch Architecture. Dotted lines represent the paths of packets addressed to a congested output.

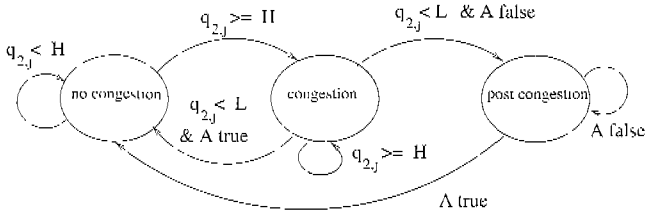

Fig. 3. State diagram of output queues. *A* represents the condition expressed by relation 2.

BSSEs has been investigated from a technological point of view and has been shown to be feasible for some units of cascading matrixes [18].

The aim is to exploit the delay lines set in the two BSSEs, each one being able to store up to $M-1$ packets by means of M delay lines. In particular, the first switch of the tandem is used to provide additional memory to limit packet loss in the second switch when any of its queues gets congested. Whereas queues of the second switch are organised as output queueing the ones of the first are dynamically managed, shared or separate, depending on the state of second stage queues. Therefore, it is necessary to properly set up this management in order to exploit the whole buffering capacity of the couple of switches.

The solution here proposed consists of sharing the storage capability at the first stage by distributing incoming packets in a round robin fashion. At the beginning, round robin operates on all queues. First stage queues are logically shared until congestion arises in the second switch. A queue is said to be congested when it contains H packets, with $H \leq M-1$. When a queue gets congested, the additional storage necessary to prevent packet loss is provided by a first stage queue that is dedicated to the congested output so as to form a unique output queue throughout the two stages. Then, during congestion only one packet per time slot is allowed to be transferred to the second stage queue and the round robin is limited only to a subset of first stage queues. Congestion ends when the queue content gets lower than a prefixed value L, $L \leq H$.

The following aspects of the tandem switch are of particular relevance and will be discussed in detail:

- management of the tandem switch queues: when and how output queues can exploit the shared memory provided by first stage switch and how to control the use of first stage queues;
- queuing algorithms for the first stage to ensure fairness and to avoid out-of-sequence.

IV. QUEUE MANAGEMENT

The reference operating environment is time-slotted, with the slot duration equal to the transmission time of a packet at the switch operating speed. Let $(1,i)$ be the i-th queue from the top of the first switch of the tandem and $q_{1,i}$ the number of packets stored in the current time slot; let $(2,j)$ be the j-th queue from the top of the second switch and $q_{2,j}$ the packets stored in the current time slot, with $i,j = 1 \ldots N$ (Fig. 2). The queues at the first stage are used to extend the storage capability offered by the second stage. This function is activated when occupancy $q_{2,j}$ of a second stage queue reaches or overcomes the prefixed value H. This causes the change of the state of the output queue as indicated in Fig. 3. When this happens, a first stage switch queue is extracted from the logical shared memory and dedicated to the congested queue. Thus, for sake of simplicity but not necessarily, let us choose $(1,j)$ as the extracted queue that provides the additional storage to enhance the buffer capacity of the overloaded output $(2,j)$, allowing only one packet per time-slot to be transferred to $(2,j)$. Due to the round robin policy used for sharing memory, all first stage queues are equivalent to this end.

Once $q_{2,j}$ gets less than L, $(1,j)$ might be reinserted in the shared memory and loaded again by the round robin cycle. However, to avoid out-of-sequence of packets, some conditions, discussed in the next section, must be verified before a queue re-enters the shared memory.

Fig. 3 shows the state diagram of an output queue considered as the virtual queue that serves a given switch output:

- no congestion: a queue remains in this state while $q_{2,j} < H$;
- congestion: this state is entered when $q_{2,j} \geq H$; one first stage queue is taken out from the round robin and dedicated as a pure output queue to the congested queue to increase its capacity;
- post-congestion: this state is entered when $q_{2,j} < L$ while the first stage queue is waiting for being re-inserted in the round robin, as a consequence of the out-of-sequence avoidance mechanism.

In order to control the transitions between these states two different techniques have been considered and described in the following.

A. Threshold Driven Control

A possible method for output queuing management is based on the definition of a threshold thr as a fraction of buffer size M: $H = thr \times (M-1)$ represents the limit value of the buffer occupancy and if H is not integer, truncation is done. $M-1-H$ is then the number of packet places available for absorbing possible multiple arrivals. When $q_{2,j} \geq H$, $(2,j)$ is said to be congested. Performance depends on H: if it is close to $M-1$, almost the whole queue is used for packet buffering but multiple arrivals can lead to packet loss; on the other hand, a low value of H absorbs well multiple arrivals but makes the output queue to get congested sooner thus increasing the overload in the first stage. The effect of the threshold position, expressed as a fraction of the buffer length, on performance is shown in Fig. 4 and values near 0.8 are recommended to minimize packet loss.

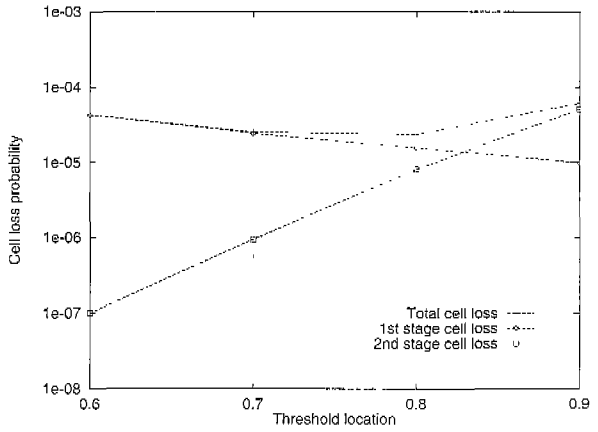

Fig. 4. Packet loss probability as a function of the threshold for a 16×16 switch with 0.8 random traffic load and buffer size equal to 12.

The congestion ends when $q_{2,j} < L$ and the related output queue enters the no congestion or the post-congestion state. This depends on whether the queue $(1, j)$ can be re-inserted in the shared memory or not in order to avoid packet out of sequence.

This mechanism requires the notification of the state of second stage queues to the first stage switch control. To this end interstage signaling should be introduced for example by applying wavelength division multiplexing of signals on the optical fibers between the first and the second stage. This implies a few modifications to the optical hardware.

B. Counters Driven Control

The previous mechanism does not fully exploit buffering capacities of the output queues since it does not keep track of the current queue occupancy but it just detect the threshold crossing. Therefore, as mentioned, packet loss may occur during interstage packet transfer because of multiple packets addressed to the same output when the place available is not enough. The mechanism here described aims at better exploiting the memories trying to perform as a pure output queue of total length $2M - 2$. It is based on the characteristic of the broadcast-and-select switch to make packets available for outputs, on their arrivals, with all possible delay between 0 and $M - 1$.

Two sets of counters are used, one couple per output link, managed by the control logic of the first stage. The first set of counters, C_{1j}, keeps track of how many packets addressed to output j are awaiting at the beginning of the current time slot in the first switch queues. C_{1j} is increased by one as a packet addressed to output j is inserted in any first stage queue, no matter if shared or separate. It is decreased by one for each packet addressed to output j forwarded to the second switch. The second set of counters, C_{2j}, keeps track of how many packets are stored in queue $(2, j)$ of the second switch at the beginning of the current time slot. It allows to compute the residual storage capacity of each second stage queue. C_{2j} is increased by one as a packet addressed to output j is forwarded from first stage and decremented by one for each time slot when it is non zero, to take into account the transmission on external links.

To avoid the interstage packet loss, the maximum number of packets to forward is limited by the residual storage capacity in $(2, j)$; this determines also the number rr_j of packets addressed to $(2, j)$ that can be stored in the shared memory of the first switch by round robin. The packets exceeding rr_j must be stored in a new dedicated queue and, as a consequence, $(2, j)$ enters the congestion state. In order to compute rr_j, let d_j be the distance, in time slot units, between the current time slot and the one when the last packet addressed to $(2, j)$ has been stored in the shared memory. We have

$$rr_j = (M - 1) - C_{2j} - C_{1j} + 1 + d_j. \tag{1}$$

For each incoming packet (1) is computed: if $rr_j > 0$ the packet is stored in the shared queue with round robin, otherwise the packet is stored in a separate queue and $(2, j)$ enters the congestion state. The congestion ends when $C_{2j} = M - 2$, that is when the queue at the second stage is no more full, and the output queue enters the no congestion or the post-congestion state. This functioning corresponds to choosing $H = L = M - 1$ and leads to packet loss performance improvement with respect to the threshold-driven control. This control technique does not need signals to be backwarded from the second stage thus the photonic complexity is not changed, providing that counters are suitably updated.

V. QUEUING ALGORITHMS FOR THE FIRST STAGE QUEUES

A. Packet Loss Balancing

All first stage queues are shared among all inputs until congestion arises at the second stage. This sharing is achieved through a top-down round robin distribution policy of incoming packets. At each time slot the round robin begins to assign packets at the first stage queues starting from the first queue not loaded in the previous time slot. Queue service, i.e., packet transfer from the first to the second stage, takes place starting from the top queue. This approach is followed to maintain packet sequence. When some queues are not available for sharing because they are dedicated to congested outputs, they are skipped by the round robin policy. If the output queue is in the post-congestion state a test is performed for its possible reinsertion in the round robin of the dedicated first stage queue at its turn.

When the queues in the round robin are all empty, the queue from which the round robin starts must be chosen. Two different approaches are considered:

- fixed, i.e., always the same queue
- variable.

With the former approach the first incoming packet is stored in the first available shared queue (not dedicated) starting from $(1, 1)$ to $(1, N)$, for instance. Accordingly, the second switch starts processing packets from input 1 to N. Simulation results reported in Figs. 5 and 6 show that this method leads to a strong asymmetry in queue loading which causes an unbalanced distribution of packet loss over all destinations. This is due to the fact that some queues stay longer than other in the post-congestion state because the number of arrivals is too low to involve those

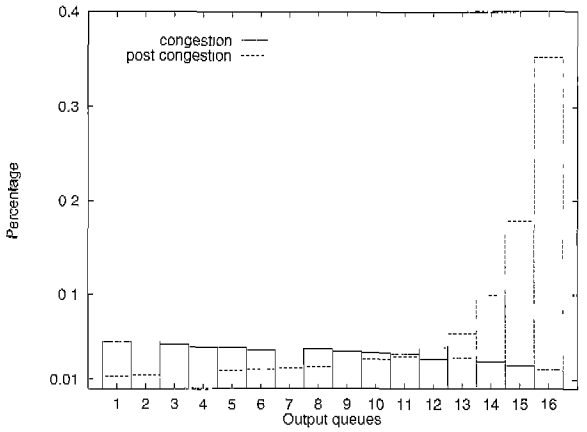

Fig. 5. Percentage time of congestion and post congestion occurred for each queue for a 16 × 16 switch with 0.8 random traffic load, buffer equal to 8 and threshold at 85% the buffer size.

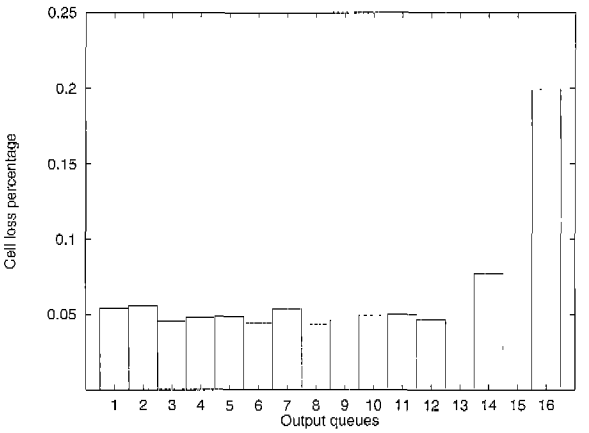

Fig. 6. Packet loss percentage for a 16 × 16 switch with 0.8 random traffic load, buffer equal to 8 and threshold at 85% the buffer size.

queues, starting from the top, or new arrivals for these post-congested queues cause the condition for re-insertion to be no longer valid.

To solve this unfairness the round-robin starting point must be moved as soon as possible. When the shared memory is empty, any not dedicated queue can be taken as first to be loaded and kept as first as long as there are packets in at least one shared queue. The second switch must be informed of the choice to accordingly process transferred packets. This method leads to a balance distribution of congestion and packet loss over all destinations (Fig. 7).

B. Out-of-sequence Avoidance

Out of sequence of packets belonging to the same virtual circuit might occur at the outputs of the first stage switch because of the round robin policy if no control is introduced. In particular, the critical point is when to re-insert a dedicated queue, let us say $(1, j)$, in the shared queue. In fact, when the round robin considers $(1, j)$, that is supposed to correspond to an output queue in a post-congestion state, the control logic must decide if it is possible to feed it with a new packet addressed to any output. The condition that must be verified is

$$q_{1,j} = q_{1,x} - 1, \tag{2}$$

where $1, x$ is the latest not congested queue that accepted a packet immediately preceding $1, j$ in the round-robin. This ensures that queue $1, j$ has been sufficiently emptied to avoid out-of-sequence otherwise the queue is not added to the shared queue. Let us refer to this condition as "*condition A*" as shown also in Fig. 3. If A is true transition to the no congestion state takes place.

VI. PERFORMANCE ANALYSIS

A. Traffic Patterns

The traffic here considered has been modeled with the two-state Markov chain usually adopted to deal with bursty traffic [19]. Burstiness b is assumed as the ratio between the average

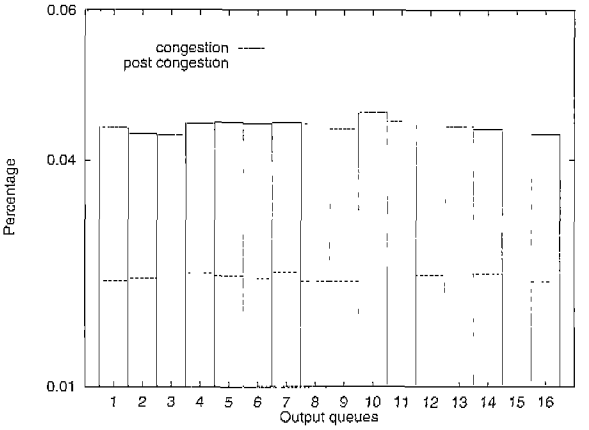

Fig. 7. Percentage time of congestion and post congestion occurred for each queue for a 16 × 16 switch with 0.8 random traffic load, buffer equal to 8 and threshold at 85% the buffer size.

burst length of the bursty traffic and the average burst length of the Bernoulli traffic with the same average load. So its value and the value of the average load completely define the bursty traffic characteristics. In particular $b = 1$ gives Bernoulli traffic. Bursty traffic can be source bursty when all packets of the burst are addressed to same destination, i.e., belong to the same virtual circuit, or network bursty, when the burst is a multiplex of packet belonging to different virtual circuits. In photonic networks as the one proposed in the environment of the European ACTS KEOPS Project, the Optical Transparent Packet Network (OTP-N), interworking units are used at the ingress of the transport photonic network to adapt the local packet formats, ATM packets for instance, with the photonic packets. This implies that traffic is smoothed and mostly can be characterised as random [20].

Traffic in a multiservice environment exhibit also point-to-multipoint characteristics. When a packet has to be broadcasted the first stage of the switch can be used to this end, being the architecture of the BSSE particularly suitable for this task. Also in this case the random traffic results can be used to obtain switch performance, by referring to a suitably increased value of average random load according to the discussion developed in [20].

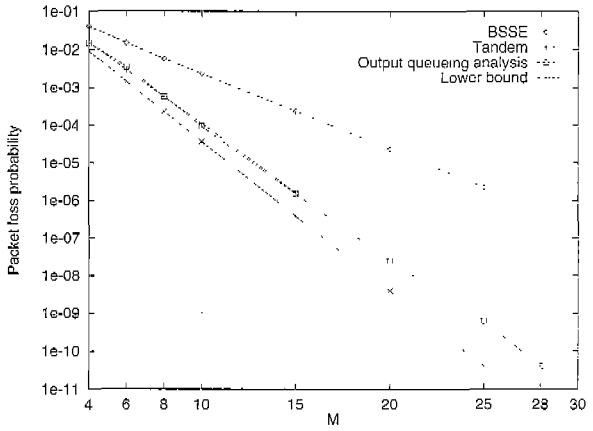

Fig. 8. Packet loss probability as a function of the number M of fiber delay lines used in each switching element for each output port with random 0.8 traffic for a 16×16 switch, comparing the single BSSE and the tandem architectures: Analysis and simulation results for the threshold driven control.

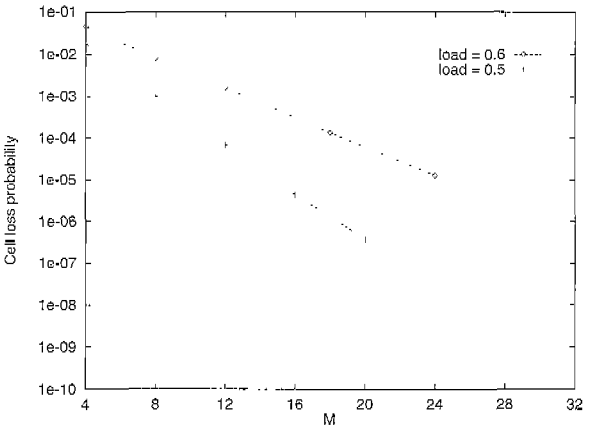

Fig. 9. Packet loss probability as a function of the number M of fibre delay lines with source bursty point-to-point traffic for a 16×16 switch: Burstiness= 1.5.

Particular attention should be given to the presence of source bursty traffic, being its impact on packet loss performance very strong. The effect of the KEOPS IWU (Inter Working Unit) on this kind of traffic is of smoothing the input burstiness when it is greater than 1 [21]. Here some results are given in order to both outline the benefits of the proposed architecture also with source bursty traffic and to find out the limitation on the average load that source bursty traffic patterns imply, when the burstiness is small.

B. Simulation Results

The performance behavior of the proposed architecture is related to output queuing performance, being the available queue length upper bounded by the sum of serialized queue lengths. Some loss in the use of available buffer locations should be accounted for, in spite of the presence of the threshold H. Performance evaluations in terms of packet loss probability have been first carried out by simulation and compared with those of the single BSSE. The evaluations have been performed for $thr = 0.85$ which simulation shows to be convenient to limit packet loss. Again, when $thr \times (M - 1)$ is not integer, truncation is done.

Our goal was to find a solution for a photonic switch to support 0.8 random traffic in the range of packet loss probability of 10^{-10}, range not achievable by a single BSSE. Fig. 8 shows the packet loss probability as a function of the number M of fibre delay lines for a 16×16 switch with 0.8 random traffic, comparing a tandem architecture with the threshold driven control with a single BSSE. It is worth noting that for a buffer size equal to 27 it can be reasonably assumed that with the tandem architecture and the buffer control mechanism described the broadband requirements are met. Also, the behaviour of the curves shows that the lower the packet loss the more convenient is the tandem. In fact, by increasing the buffer size of one unit in the BSSE, every tandem switch queue is increased of one unit as well, which in its turn means to increase of 2 units the whole output queue. Thus, if we could use about 50 fibre delay lines for each queue in a single switch, the single BSSE would be enough, but since the greatest number of fibre delay lines for each output expected in the near future is 32 [20], a tandem architecture shall be used to meet a packet loss probability of 10^{-10} with only 27 delay lines.

Being the counter driven control a solution that prevent loss in the second stage of the tandem, its packet loss performance are expected to be related to those of an output queue with length equal to the sum of the buffer capacity that can be exploited in the switches of the tandem, that is its performance are expected to achieve the lower bound. This has been proved by simulation thus showing the possibility to fully exploit the serialized buffer capacity.

Fig. 9 shows packet loss performance in the presence of source bursty traffic with burstiness 1.5. It can be seen that limitation on the average load is necessary to stay within buffer constraints. In any case the approach adopted allows up to 0.5 average load with 32 delay lines at packet loss rate 10^{-10}.

C. Application of the Output Queueing Analysis

Being the behavior of the queues in the congestion state related to that of a whole queue obtained by the serialization of corresponding queues in the two switches, packet loss performance are strictly related to those of a switch with output queueing. A lower bound on packet loss is so represented by analytical evaluations performed for a queue length equal to the sum of the serialized two queues, that is $2M - 2$ buffer locations.

If we indicate with $\Phi(q)$ the packet loss probability obtained for a queue length q by means of the classical output queueing with cut-through model [10] for a given load and switch dimension, $\Phi(2M - 2)$ represents the lower bound shown in Fig. 8. The difference from simulation is related to the presence of the threshold thr does not allow a full exploitation of the total serialized buffer capacity.

An approximated evaluation of the packet loss behavior is so expected to be given by the sum of buffer length $M-1$ at the first switch and the value of the threshold in the second switch, that is $\Phi(M - 1 + \lfloor thr \times (M - 1) \rfloor)$. Results are given in Fig. 8 and show a perfect agreement with simulation results. So the tandem switch achieves output queuing packet loss performance with

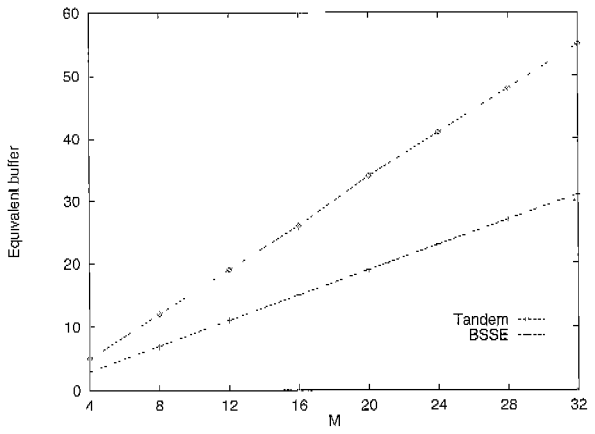

Fig. 10. Equivalent buffer capacity realized by the tandem architecture as a function of the number M of fibre delay lines used in each switching element for each output port: Comparison with the BSSE.

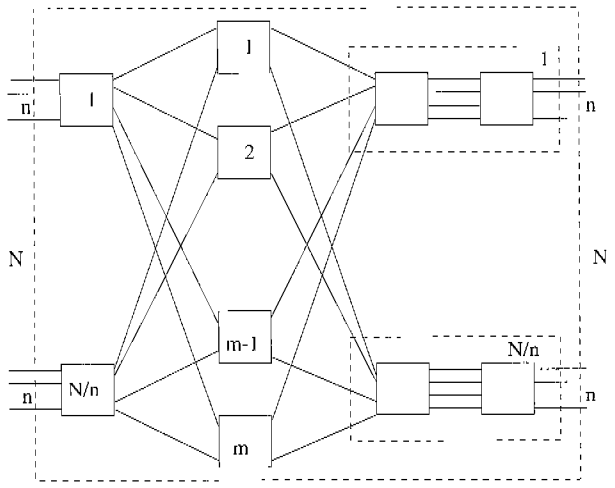

Fig. 11. The Clos multistage architecture.

equivalent queue length $Q = M - 1 + \lfloor thr \times (M - 1) \rfloor$. This also means that packet loss performance of the tandem switch can be exactly calculated by an analytical model.

In Fig. 10 the value of the equivalent output queue realized by means of the tandem queue is plotted as a function of the number of delay lines in each switch for $thr = 0.85$. It can be seen the linear growth of the equivalent buffer capacity as a function of M that allows to obtain queue lengths up to about 50 packets with moderate hardware increase.

VII. DESIGN OF A LARGE SWITCH

A large monolithic $N \times N$ output queuing switch is practically unfeasible when N overcomes some tens so multistage architectures can be a solution [18]. Different options are available in the organization of small switching matrixes into a multistage architecture and are widely discussed in[11], [22], [23]. In particular performance are sensibly influenced by the topology of the interconnection network.

A flexible choice is represented by a three-stage Clos multipath architecture [24] (Fig. 11). If the number of SEs at the intermediate stage is greater than that at the first stage we have a

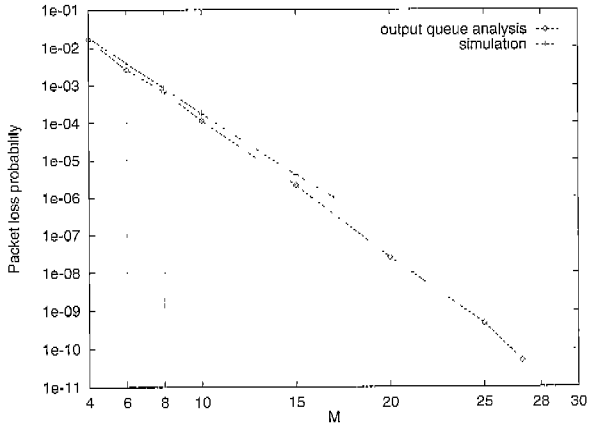

Fig. 12. Packet loss probability as a function of the number M of fibre delay lines: Comparison of analysis with random traffic and simulation with multiplexed bursty traffic, burstiness= 1.5, for a 32×32 switch with $2:1$ internal expansion, load 0.8.

"Clos architecture with expansion" ($m > N/n$). The number of I/O connections and the number of SEs at each stage are design freedom degrees to be chosen in a cost/performance trade-off. When expansion is introduced packet loss is shifted from the first and second stage to the third one, which behaves as a concentration stage and the whole switch acts as an output queueing switch. Buffer requirements for this stage have been shown to overcome the today photonic technology [20] and some improved buffer management technique is necessary. For example the tandem solution can solve the buffer dimensioning problem at this third stage: in Fig. 12 performance, calculated as $\Phi(M - 1 + \lfloor thr \times (M - 1) \rfloor)$ are shown for the third stage of a 32×32 switch with 1:2 internal expansion at 0.8 random load. Switching elements at the third stage are thus 32×16 with 0.4 load. The value of thr is set to 0.85. The results obtained are compared with simulation results with burstiness = 1.5, to take into account the effects of cascading stages on traffic pattern [20], and the resulting buffer requirements are shown to be within technological limitations. Analysis slightly differs from simulation because of the burstiness taken into account by the latter. The irregularity of the curve relative to the analysis is due to the effect of truncation applied to obtain the equivalent queue length Q.

VIII. CONCLUSIONS

In this paper we have presented the tandem architecture based on the broadcast and select photonic switches in order to get better packet loss probability performance by improving the management of the output optical buffers. These improvements are possible by applying suitable control procedures for routing and storing packets in the switches of the tandem.

The results here presented show a significant improvement with respect to the single switching element, that has been proved to be strictly related to output queuing performance. For the threshold driven control, simulation and analysis are in perfect agreement if a buffer size is assumed given by the sum of the buffer size at the first switch and the portion defined by the threshold at the second switch. The counter-driven control has

proved to reach the lower bound on packet loss performance. An example of application of the proposed architecture has also been given as a concentration stage of a multistage architecture for large switches, showing the feasibility of the solution for actual implementations.

ACKNOWLEDGMENT

The authors wish to thank Miss Gisella Montosi and Mr. Paolo Vignoli for their help in the development of the simulation programs.

REFERENCES

[1] ITU, "World telecommunication progress report 1996/97," Feb. 1997.
[2] P. Gambini, "State of the art of photonic packet switched networks," in *Proc. of International Workshop on Photonic Networks & Technology*, Sept. 1996, Lerici, Italy.
[3] F. Masetti *et al.*, "High speed, high capacity ATM optical switches for future telecommunication transport networks," *IEEE J. Select. Area Commun.*, vol. 14, no. 5, June 1996.
[4] C. Guillemot *et al.*, "Transparent optical packet switching: the european ACTS KEOPS project approach," *IEEE-OSA J. of Lightwave Technology*, Invited paper, vol. 16, no. 12, pp. 2117–2134, Dec. 1998.
[5] M. Renaud, F. Masetti, C. Guillemot, and B. Bostica, "Network and systems concepts for transparent optical packet switching," *IEEE Commun. Mag.*, Apr. 1997.
[6] F. Callegati, M. Casoni, C. Raffaelli, and B. Bostica, "Packet optical networks for high-speed TCP-IP backbones," *IEEE Commun. Mag.*, pp. 124–129, Jan. 1999.
[7] D. Chiaroni *et al.*, "Rack-mounted 2.5 Gbit/s ATM photonic switch demonstrator," in *Proc. of ECOC'93*, Montreux, CH, post-deadline paper ThP12.7.
[8] D. Chiaroni *et al.*, "A novel photonic architecture for high capacity ATM switching applications," in *Proc. of Photonics in Switching '95*. Mar. 1995, Salt Lake City, UT, USA, paper P.Th.C3.
[9] F. Masetti *et al.*, "Fiber elay lines optical buffer fors ATM photonic switching application," in *Proc. of INFOCOM'93*, 1993, S. Francisco, CA, USA, pp. 935–942.
[10] M. J. Karol, M. G. Hluchyj, and S. P. Morgan, "Input versus output queueing on a space-division packet switch," *IEEE Trans. Commun.*, vol. COM-35, no. 12, pp. 1347–1356, Dec. 1987.
[11] M. Casoni, G. Corazza, J. B. Jacob, F. Masetti, and C. Raffaelli, "Clos architecture for the design of large photonic ATM switches", in *Proc. of EFOC & N'94*, June 1994, Heidelberg, Germany, pp. 106–110.
[12] D. K. Hunter, M. C. Chia, and I. Andonovic, "Buffering in optical packet switchies", *IEEE-OSA J. Lightwave Technology*, vol. 16, no. 12, pp. 2081-2094, Dec. 1998.
[13] F. Callegati *et al.*, "Performance evaluation of a new photonic ATM switching architecture based on WDM," in *Proc. of Australian Telecommunication Networks & Application Conference*, Dec. 1995, Sydney, Australia, pp. 387–392.
[14] P. Doussiere *et al.*, "1550 nm polarisation independent DBR gain clamped SOA with high dynamic input power range," in *Proc. of ECOC'96 conference*, paper WeD 2.4, Sept. 1996, Oslo, Norway.
[15] G. Soulage *et al.*, "Clamped-gain SOA gates as multiwavelength space switches," in *Tech. Digets OFC'95*, Feb. 1995, San Diego, CA, USA, Paper TuD1
[16] D. Chiaroni *et al.*, "Wavelength channel selector with subnanometric resolution and subnanosecond switching time," in *Proc. of ECOC'94*, Florence, Italy, pp.593–596.
[17] T. Durhuus *et al.*, "Penalty free all-optical wavelength conversion by SOA1's in mach-zehnder configuration," in *Proc. of ECOC'93*, Montreux, Switzerland, paper TuC5.2, vol. 2, pp. 129–132.
[18] D. Chiaroni *et al.*, "Theoretical feasibility analysis of a 256x256 ATM optical switch for broadband applications," in *Proc. of ECOC '93*, Sep. 1993, Montreaux, Switzerland, pp. 485–488.
[19] Y. Xiong, G. Petit, and H. Bruneel, "Performance study of an ATM self-routing multistage switch with bursty traffic: Simulation and analytical approximation," *European Trans. Telecom*, vol. 4, no. 4, July-August 1993.
[20] F. Callegati *et al.*, "Architecture and performance of a broadcast and select photonic switch," *Optical Fiber Technology*, Academic Press, Invited paper, pp. 266–284, July 1998.
[21] M. Calisti and F. Callegati, "Traffic models for an optical transparent packet network," in *Proc of NOC '97*, 17-19 June 1997, Antwerp, Belgium, pp. 55–62.
[22] M. Casoni, C. Raffaelli *et al.*, "Performance analysis of a 256x256 ATM photonic modular switching fabric," in *Proc. of 1st IEEE BSS*, 19-21 Apr. 1995. Poznan, Poland, pp. 101–107.
[23] J. B. Jacob, M. Casoni, G. Corazza, F. Masetti, P. Parmentier, and C. Raffaelli, "System design and evaluation of a large modular photonic ATM switch," *European Trans. Telecom.*, vol. 7, no. 6, pp. 565–573, Nov. 1996.
[24] C. Clos, "A study of non-blocking switching networks," *Bell Systems Technical Journal*, pp. 406–424, Mar. 1953.

Maurizio Casoni graduated in Electrical Engineering from the University of Bologna in 1991 with honors, with a grant supported by Telecom Italia and received the Ph.D. degree in EE also from the University of Bologna, in 1995. In 1995 he was with the Washington University in St.Louis, MO, as a research fellow. Currently he is Research Associate at the Engineering Sciences Dept. of the University of Modena and Reggio Emilia. He has studied ATM broadband switching architectures, analytical models for shared buffer switches, Clos architecture for the design of large photonic switches, congestion control mechanisms based on selective packet discard. He was involved in the ACTS KEOPS project of the EU on the photonic transport of information. Currently he is involved in national projects on radiomobile systems for the support of multimedia applications and techniques for providing QoS in the Internet.

Carla Raffaelli received her Electronic Engineering degree from the University of Bologna, Italy, in 1985 and the Ph.D. Degree in Electronic Engineering and Computer Science in 1990. Since 1985 she has been with the Department of Electronics, Computer Science and Systems of the University of Bologna, where she became a research associate in 1990. Her research interests are in the field of broadband communication, protocols and modeling. In 1990-94 she has been involved in research on ATM networks and switching in the framework of the Telecommunication Project supported by the Italian National Research Council (C.N.R.). She was involved in the project KEOPS of the European Community on photonic transport of information and in other national research programs on high speed wireless local area networks. She is now involved in national projects on techniques for providing QoS in the Internet.