EUROPEAN TRANSACTIONS ON TELECOMMUNICATIONS
Eur. Trans. Telecomms. 2009; **20**:734–745
Published online 1 June 2009 in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/ett.1365

Optical Communications

Guaranteeing packet order in IBWR optical packet switches with parallel iterative schedulers

M. Rodelgo-Lacruz[1], P. Pavón-Mariño[2], F. J. González-Castaño[1,3]*, J. García-Haro[2], C. López-Bravo[3], J. Veiga-Gontán[2], F. Gil-Castiñeira[3] and C. Raffaelli[4]

[1]*Galician Research and Development Center in Advanced Telecommunications (GRADIANT), Spain*
[2]*Universidad Politécnica de Cartagena, Spain*
[3]*Universidad de Vigo, Spain*
[4]*Università di Bologna, Italy*

SUMMARY

The input-buffered wavelength-routed (IBWR) switch is a scalable switch fabric for optical packet switching (OPS) networks. In synchronous operation, when optical packets are of a fixed duration and aligned at switch inputs, the scheduling of this architecture can be characterised by a type of bipartite graph matching problem. This challenges the design of feasible algorithms in terms of implementation complexity and response time. A previous work presented and evaluated the insistent parallel desynchronized block matching (I-PDBM) algorithm for the IBWR switch. I-PDBM is a parallel iterative scheduler with a good performance and a simple hardware implementation. However, the algorithm does not maintain the packet sequence. In this paper, we present the I-PDBM algorithm with packet ordering (OI-PDBM), which prevents mis-sequencing and behaves as I-PDBM in terms of delay, buffer requirements and convergence speed. Copyright © 2009 John Wiley & Sons, Ltd.

1. INTRODUCTION

In the optical packet switching (OPS) paradigm of wavelength division multiplexing (WDM), packet payloads stay in the optical domain. OPS offers high bandwidth efficiency due to statistical multiplexing, but it is well known that packet granularity and optical buffering impose extreme constraints on photonic witching, incurring unacceptable hardware costs with state-of-the-art technology.

In this paper, we focus on synchronous slotted OPS, which specifies a fixed packet size (slot length) and packet alignment with slot boundaries at the input ports. Packet alignment requires optical synchronising stages, which increases switch cost. However, its better contention behaviour has encouraged the study of this alternative. The European DAVID project [1] selected synchronous slotted OPS with slot lengths of $\sim$1 μs as a medium-term alternative for the WDM backbone network.

In WDM OPS networks, permanent end-to-end connections (Optical Packet Paths, OPPs) are provisioned to follow a fixed sequence of hops from ingress to egress nodes. The scattered wavelength path (SCWP) operational mode [2] means that the packet transmission wavelength in each hop is not fixed. This provides extra freedom to the switch schedulers in packet wavelength selection, boosting the statistical multiplexing effect. Therefore, SCWP achieves a high throughput and a low packet delay in OPS networks, thus also lowering optical buffering requirements [3, 4]. By nature, this effect is particularly relevant in the dense wavelength division multiplexing (DWDM) scenario: the higher the number of wavelengths per fibre, the more intense the multiplexing gain. The research in this paper assumes a SCWP OPS network.

* Correspondence to: F. J. González-Castaño, Galician Research and Development Center in Advanced Telecommunications (GRADIANT), Spain.
E-mail: javier@det.uvigo.es

Received 4 November 2007
Revised 8 October 2008
Accepted 23 March 2009

Although packet order is not strictly necessary in Internet routers, common practice dictates that switches should not disorder packets unless strictly necessary. As an example, packet mis-sequence can cause undesirable effects which affect the congestion control schemes of TCP versions [5, 6]. The effect of OPS packet re-ordering on TCP behaviour has been considered in Reference [7]. In a backbone network, with a connection-oriented traffic demand, it would be enough to guarantee packet order among packets within the same traffic flow. Nevertheless, in electronic switching, it is common to follow a suboptimal approach, and keep packet sequence among all packets that enter the switch through the same input port, and leave the switch through the same output port (whether they belong to the same traffic flow or not).

In an OPS backbone network, which is intended to carry future Internet traffic, it is desirable to preserve the end-to-end packet sequence. Electronic re-sequencing stages at the egress nodes should be avoided, as they would need very large memories due to the high speed of the optical links. Assuming this, the ingress nodes and the interconnection nodes must enforce packet sequence in each hop across the network. Thus, in principle, packet order information should be available for switching nodes. However, adding sequence information into the packet header is not desirable due to the performance degradation as a consequence of header enlargement. Alternatively, it is possible to design an ordering criterion based on packet arrival time and packet arrival wavelength. Packet arrival time is not enough for packet ordering in SCWP networks, because several packets belonging to the same OPP may be simultaneously transmitted using different wavelengths. When these simultaneous packets arrive at the next node in the path, the switch scheduler requires extra information to know their original order and preserve it in the delay assignment and in the output wavelength allocation.

The Wavelength Switched Packet Network (WASPNET) packet sequencing criterion for OPS networks proposed in Reference [8] consists of transmitting simultaneous packets in consecutive wavelengths, starting from the lowest wavelength, so that lower order packets have lower wavelengths assigned. However, this criterion unbalances the wavelengths in the fibres (lower order wavelengths are used more frequently than higher order ones) and degrades switch performance [9]. In this paper, we adopt the round-robin criterion presented in Reference [9] that balances the average use of wavelengths per fibre. If pck_i and pck_{i+1} are consecutive packets, transmitted in a WDM link in time slots $t(\text{pck}_i)$ and $t(\text{pck}_{i+1})$, and wavelengths $\lambda(\text{pck}_i)$ and $\lambda(\text{pck}_{i+1})$, respectively, the round-robin ordering criterion specifies that: (a) $t(\text{pck}_{i+1}) > t(\text{pck}_i)$, and (b) $\lambda(\text{pck}_{i+1}) = (\lambda(\text{pck}_i) + 1) \bmod n$, where n is the number of wavelengths for the fibre under consideration. The result, as shown in Figure 1, is an exact round-robin packet spread across the wavelengths, for any traffic pattern in the fibre. The criterion requires each node to remember the wavelength of the last packet received/transmitted in the sequence across consecutive time slots. As a consequence, the implementations of this functionality require two sets of round-robin pointers to track packet sequence: one round-robin pointer per input fibre, tracking the wavelength of the next packet in the input traffic sequence, and one round-robin pointer per output fibre, determining the output wavelength of the next packet to be transmitted.

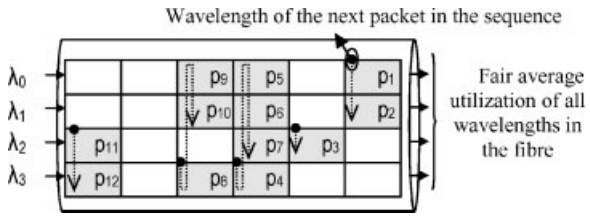

Fig. 1. Round-robin wavelength sequence criterion, fibre with four wavelengths $\lambda_0,...,\lambda_3$.

This paper focuses on the input-buffered wavelength-routed (IBWR) switch architecture proposed in Reference [10]. Figure 2 shows the WDM adaptation of this architecture. The switch has N input/output fibres, and n wavelengths per fibre. It has a buffering section followed by a non-blocking switching section. The buffering section consists of nN tunable wavelength converters (TWC) with a tuning range $\lambda_0 ... \lambda_{K-1}$, $K = \max(nN, M)$ and two $K \times K$ arrayed waveguide gratings (AWGs). We denote as M the number of delay lines, of sizes from 0 to $M-1$ slots, interconnecting the two AWGs. Due to AWG symmetry

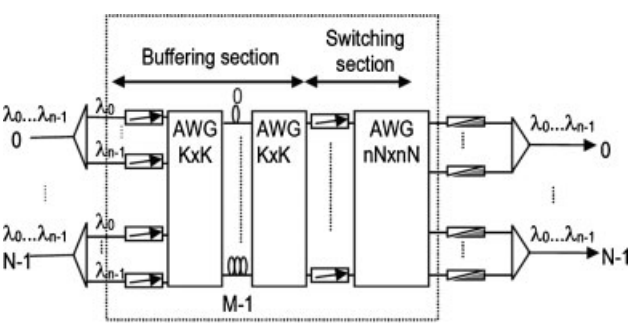

Fig. 2. Input-buffered wavelength-routed switch (IBWR).

DOI: 10.1002/ett

[11], a packet arriving at port i leaves the buffering section at port i, regardless of the selected delay. The wavelength conversion at the buffering section determines the delay line for the packet. The switching section is composed of nN TWCs followed by an $nN \times nN$ AWG. The wavelength conversion at the switching section determines the output fibre/wavelength for the packet.

The IBWR switch scheduler assigns packet delays and packet output wavelengths. These two tasks are independent.

Packet delay assignment. Current optical switches employ fibre delay lines (FDLs) due to the lack of optical RAMs. In IBWR switches, delays are assigned at packet arrivals. The scheduler discards a packet if it cannot assign a delay fulfilling two contention conditions: (a) *output fibre contention*: at most n packets can reach any output fibre in a given slot, (b) *input port contention*: the packets that arrive at the same ith input port (same fibre and wavelength) in different time slots cannot leave the switch in the same time slot. Otherwise they would collide at the ith TWC of the switching section, which can only handle one packet at a time.

Remark: Other OPS architectures, with higher hardware costs and less scalability than IBWR, emulate output buffering (OB) [10, 12] (the only factor limiting packet delay assignment is the unavoidable output fibre contention).

Output wavelength assignment. The scheduler assigns output wavelengths to the packets when they leave the switch, according to the round-robin criterion.

Previous work has characterised IBWR delay assignment as a bipartite graph matching process [4]. At every slot, the scheduler seeks a feasible assignment maximising the number of packet delay assignments (i.e. minimising packet losses). If there are several alternatives, it minimises average packet delay. The *sequential* IBWR scheduler for the SCWP mode in Reference [8] was conceived for testing purposes, for a first performance evaluation of the SCWP IBWR architecture. Its algorithm performs a sequential check of all input ports whose response time depends on switch size, which makes it impossible to fulfil the time constraints (for $\sim$1 μs slots) in practical electronic implementations, even for moderate switch sizes. Conversely, our proposal is parallel, as in the virtual output queuing (VOQ) schedulers.

The rest of this paper is organised as follows. In Section 2, we briefly describe the previous insistent parallel desynchronized block matching (I-PDBM) scheduler. In Section 3, we present the new I-PDBM algorithm with packet ordering (OI-PDBM). In Section 4, we evaluate its performance and discuss simulation results. Section 5 concludes the paper.

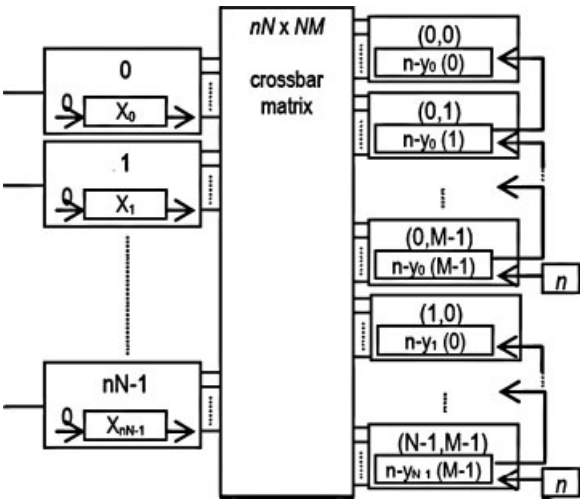

Fig. 3. Electronic implementation scheme for the I-PDBM scheduler.

2. PREVIOUS WORK: I-PDBM SCHEDULER

In Reference [13] we presented I-PDBM, an enhanced version of the parallel desynchronized block matching (PDBM) algorithm [14]. Figure 3 illustrates an electronic implementation of I-PDBM. The nN input modules (one per input fibre wavelength) are interconnected with NM output modules (one per output fibre and delay line) by three signals. The *request* and *accept* signals, from input to output modules, and the *grant* signals, from output to input modules.

Input module i, $i = 0,...,nN-1$, keeps an input TWC availability state vector $\overline{X}_i(t)$, $t = 0,...,M-1$. Component $\overline{X}_i(t)$ is 1 if a packet is scheduled to leave the buffering section at the ith port in t slots (0 otherwise). At every slot the state vector is shifted: $\overline{X}_i(t-1) = \overline{X}_i(t)$ and $\overline{X}_i(M\text{-}1) = 0$ to reflect FDL propagation after each slot.

Output module (j, t), $j = 0,...,N-1$, $t = 0,...,M-1$, keeps two registers and a flag. (a) A delay availability register of $\log_2(n)$ bits, storing the value $n - y_{jt}$. Variable y_{jt} denotes the number of packets already scheduled for output fibre j that will leave the switch in t time slots. (b) A *grant pointer* FG_{jt}, of $\log_2(N)$ bits, which indicates the first input fibre to scan. (c) An alternating bit CW_{jt} indicating the search direction. At the end of every slot, packet propagation along the FDLs is manifested by transferring the delay availability register in module (j, t) to module $(j, t-1)$, $j = 0,...,N-1$, $t = 1,...,M-1$. Also, modules $(j, M-1)$, $j = 0,...,N-1$, reset the availability registers to n.

DOI: 10.1002/ett

At each input fibre controller, an arrivals round-robin pointer WG_f, $f=0,\ldots,N-1$, indicates the wavelength which carries the first packet in input fibre f at the current time slot, as dictated by the round-robin ordering criterion.

At system initialisation, $\overline{X}_i(t)$, y_{jt}, WG_f and CW_{jt} are set to 0. It is of interest that input modules which do not receive any grant from module (j, t) may receive a grant from any other output module (j, t'), $t' \neq t$. If grant pointers associated with the same output fibre had the same position (synchronised), closer input modules would receive more than one grant, while further input modules would receive no grant in this iteration. To minimise this *grant block overlapping effect* [14] the FG_{jt} grant pointers are desynchronised during system start-up, and after that, pointer desynchronising is maintained by simultaneously incrementing (modulo N) the positions of all pointers every two time slots. During initialisation, the positions of the M FG_{jt} grant pointers associated with the same output fibre are spread across the N input fibres, such that the minimum distance (modulo N) between two nodes is maximised.

$$\begin{aligned} FG(f,0) &= 0 \\ FG(f,t) &= FG(f,\ t-1) \\ &\quad + \min\left(1,\ \left\lfloor \frac{N}{M} \right\rfloor\right) \quad \begin{array}{l} \forall f = 0 \ldots N-1 \\ \forall t = 1 \ldots M-1 \end{array} \end{aligned}$$

According to the well-known iSLIP algorithm [15], the iterations of the I-PDBM algorithm consist of two steps (*request* and *grant*) and each step is executed in parallel by all the input/output modules of the scheduler. Nevertheless, the accept phase is performed just once at the end of the algorithm:

Step 1. Request: Each input module i with a packet destined to output fibre j sends a *request* signal to every output module of fibre j whose associated delay satisfies that (a) it does not violate input contention, and (b) if this is not the first iteration of the algorithm, and this packet was already granted a delay p, no request signals must be sent to longer delays other than the one already granted. In other words, the input module sends *request* signals to output modules (j, t) such that $\overline{X}_i(t) = 0$ and $t \leq p$, where p is the shortest granted delay in previous iterations.

Step 2. Grant: Each (j, t) output module scans the *request* signals from the input modules, starting with the input module indicated by its pointers FG_{jt} (grant fibre) and WG_f (arrivals wavelength). The sequence of scanned fibres proceeds in a clockwise or counter-clockwise sense, according to the alternating bit CW_{jt}. The first $n - y_{jt}$ scanned *request* signals are acknowledged, and a grant signal is sent to the associated input module. Note that this step guarantees that the output contention constraint is fulfilled.

End of the algorithm. Accept: During each time slot, a packet is discarded if its input module does not receive any grants in the last request–grant iteration. Otherwise, the shortest granted delay t is assigned to the packet that is present at input i, and an accept signal is sent to the accepted output module. Vector $\overline{X}_i(t)$ and variable y_{jt} are updated and shifted as described above to consider the allocation and the propagation of the packets in the delay lines. The WG_f round-robin arrivals pointers are increased by the number of received packets at fibre f in the current slot (modulo n), as dictated by the round-robin ordering criterion. The FG_{jt} grant pointers are incremented by one (module N) every *two* time slots to keep arbiter desynchronisation. The CW_{jt} bits are negated to alternate *request* scanning directions during each time slot, to enforce fairness in the case of non-uniform packet arrivals.

In I-PDBM, grants are provisional, until the accept stage takes place. Through request–grant iterations, each granted input module stops requesting delays higher than the best delay granted in previous iterations, but it keeps the request signal active for better ones. Since the pointers do not change until the accept step at the end of the slot is reached, each granted input module that maintains an active request signal is *granted again*. By stopping higher delay requests, it releases some wavelengths that can be granted to other modules. Subsequent iterations may increase the number of packet assignments and further reduce the delay of previously assigned packets.

I-PDBM converges in $\min(M, nN)$ iterations (independently of switch size), its performance is good in terms of delay and buffer requirements and its hardware is simple [13]. In a realistic backbone WDM topology switch size N is small and, as we can see in Section 4, if the number of wavelengths grows, the number of required delays and, consequently, the number output modules M is also small. For example, a switch of size $N=4$ and $n=64$ wavelengths just requires $M=2$ buffers to achieve a negligible packet loss probability. Thus, the scheduler has $nN=256$ input modules and just $NM=8$ output modules (256×8). Although I-PDBM has not been implemented yet, the algorithm steps are very similar to other parallel iterative maximal size matching schedulers with decision times of a few nanoseconds for 256×256 ports [16, 17], so we can expect that I-PDBM can be easily implemented for a decision time of $\sim$1 μs. However, its main drawback is packet mis-sequence.

DOI: 10.1002/ett

3. I-PDBM WITH PACKET ORDERING (OI-PDBM) ALGORITHM

In order to avoid out-of-sequence packets, two conditions have to be met:

(a) The delay line assigned to a packet must be high enough to guarantee that packet transmission takes place at the same time as or later than the transmission of the precedent packet in the same input–output fibre pair.
(b) After the previous condition, if two packets in the same input–output fibre pair are scheduled to leave the switch simultaneously, then the order in the transmission wavelength allocation should be maintained: a precedent wavelength following the round-robin order has to be assigned to the precedent packet.

The IBWR architecture does not impose any constraint on output wavelength allocation. Output wavelength allocation is identical to that of I-PDBM. The decision on the transmission wavelength of a packet occurs during the time slot in which the packet leaves the switch, not during the time slot in which the packet arrives at the switch. When a packet is assigned to a delay, it is also assigned to an increasing wavelength priority (the number of packets already scheduled for the given output fibre in t time slots). If two or more incoming packets destined to the same output fibre that arrived within the same time slot are assigned to the same delay, priority is assigned following the input wavelength priority order. At the moment of packet transmission, the transmission wavelength is easily calculated from the packet priority and the round-robin transmission wavelength pointer. This simple process can be easily implemented in hardware.

In conclusion, the mechanism implemented in the OI-PDBM algorithm only enforces condition (a), constraining the delay assigned to a packet. To do that, two different cases have to be considered among the arriving packets: (a) packets whose precedent packet arrived in a previous time slot, (b) packets whose precedent packet arrived in the same time slot.

For the first type of packets, packet ordering is easy to implement, by adding an additional state vector $\overline{z_{jt}}(i)$, $i = 0,...,N$, to each output module (j, t). Component $\overline{z_{jt}}(i)$ takes a value of 1 if a packet that arrived at input fibre i destined to output fibre j will leave the switch later than t time slots (0 otherwise) and each output module (j, t) ignores all signals from input fibre i if $\overline{z_{jt}}(i) = 1$. After a time slot, the registers are updated: if a packet from input fibre i destined to output fibre j is scheduled to leave the switch after t time slots, then $\overline{z_{jt'}}(i) = 1\ \forall t' < t$. Next, the registers are shifted to reflect packet propagation through delay lines: $\overline{z_{jt-1}} = \overline{z_{jt}}(i)$ and $\overline{z_{jM-1}}(i) = \overline{0}$. Figure 4 shows the state vectors of every input port of input fibre fi ($n = 4$): $\overline{x_{fi,\forall\lambda}}(t)$,

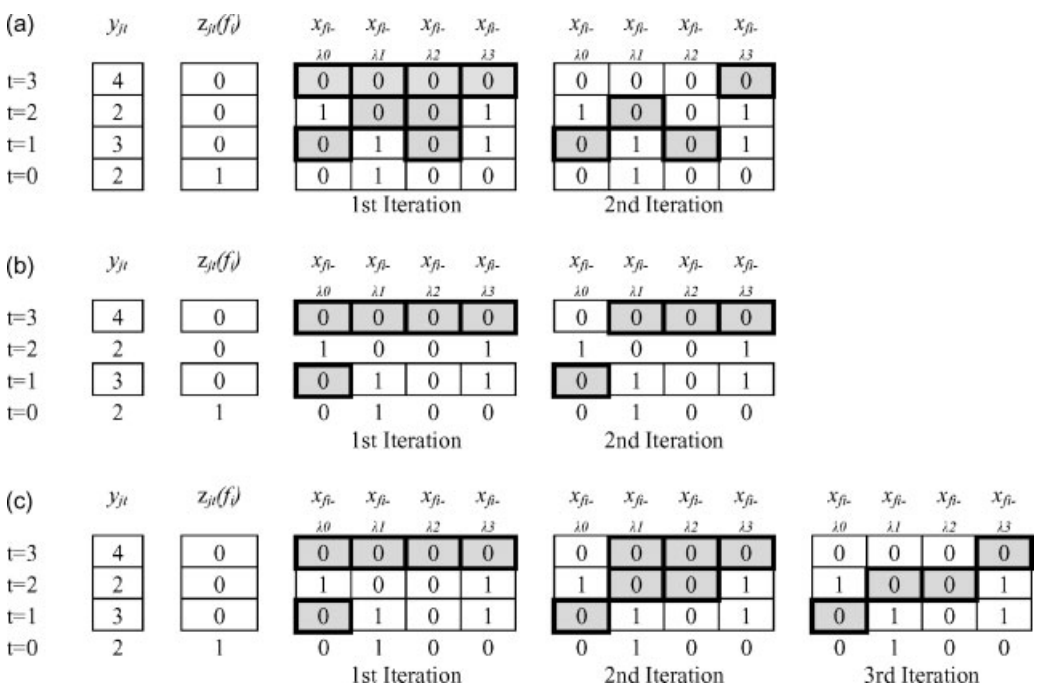

Fig. 4. Assignment examples: (a) I-PDBM with $\overline{z_{jt}}(i)$ restrictions, (b) first approximation of OI-PDBM and (c) OI-PDBM.

DOI: 10.1002/ett

the state vectors of output modules of output fibre j: $\overline{z_{jt}}$ (fi) and $\overline{y_{jt}}$, and some assignment examples (grants are represented by shaded boxes) if all the incoming packets of input fibre i are destined to output fibre j and λ_0 is the highest priority wavelength in that time slot. Figure 4(a) shows an I-PDBM assignment considering $\overline{z_{jt'}}(i)$ constraints if a packet from input fibre i to output fibre j has been scheduled to leave the switch in the next time slot. Input signals to output module (j, 0) are not disabled to avoid packet mis-sequencing. Nevertheless, the packet in wavelength λ_2 is scheduled to leave the switch before the higher priority packet in λ_1, since this schema cannot prevent disorder among packets that arrive in the same time slot.

Guaranteeing packet order between two simultaneous packets is a more complex task. It can be fulfilled if an output module only grants a delay b to an input module if the previous input modules in the same fibre (precedent in the round-robin order of arrivals) have requested the same delay. This is because, if a packet receives a grant, the round-robin precedent packets in the same input fibre are granted at least the same delay b (and maybe a shorter one by another output module). Nevertheless, the condition is too restrictive. Precedent input ports which do not have a packet to the module output fibre do not constrain the granting process in this output module. The same happens for precedent input ports which are already granted a shorter delay $b' < b$. Figure 4(b) shows an example. The assignment preserves packet order but it is clearly suboptimal. The packets in λ_1 and λ_2 are scheduled for delay $t = 3$ instead of $t = 2$, to avoid overtaking the packet in λ_0. Nevertheless, the latter has been scheduled to $t = 1$ and this restriction is unnecessary.

To enhance the mechanism, we add a signal *allow* from the input to the output modules. An input module activates the *allow* signal to indicate the output modules that they can ignore a packet as a possible source of packet disorder. These are (a) the output modules for output fibres they do not have a packet for, (b) the output modules with a delay longer than the best delay already granted in previous iterations. Therefore, the output modules safely grant input module requests only if the precedent input modules of the same input fibre have activated the *allow* signal (and thus were not granted) or the *request* signal (which means they were also granted). Figure 4(c) shows how this mechanism achieves optimal matching. In the second iteration, the input module of wavelength λ_0 receives a *grant* from delay $t = 1$ and activates the *allow* signal for delay $t = 2$. Then, output module (j, 2) can *grant* packets in λ_1 and λ_2, given that it does not cause mis-sequencing.

In I-PDBM, the input modules continue requesting the best granted delay and do not accept it until the last iteration, since it is assured that, if they request a delay granted in the previous iteration, it will be granted again. Nevertheless, in OI-PDBM, an input module with a granted delay activates the *allow* signal so that other input modules of the same input fibre can receive grants. Granting delays to these input modules may consume available wavelengths assigned in previous iterations to other input modules (*ungranted* modules). For example, granting input ports of wavelengths λ_1 and λ_2 in the second iteration of Figure 4(c) may consume wavelengths that were assigned to fibres with lower priority in the first iteration. This effect entails two drawbacks. First, algorithm convergence is slowed down by a cascade effect of ungranted modules that deactivate the *allow* signal and may produce new ungranted modules. Second, when an input module is ungranted, its *allow* signals remain active in the iteration, meaning that in intermediate iterations granted delays can temporarily cause packet mis-sequence. Although the algorithm converges to a solution without packet disorder, if there is an imposed bound on the number of iterations before convergence is guaranteed, packet mis-sequence may result.

We avoid this harmful effect as follows: at the beginning of an iteration, output modules keep active the grants for input modules that hold the *request* signal and deactivate the grants for input modules that activated the *allow* signal. The new available grants in the current iteration (to the input modules which deactivated grant signals) equal the number of available wavelengths in the fibre minus the number of activated grants.

3.1. OI-PDBM algorithm

The changes regarding I- PDBM are:

Step 1. Request: Each input module i with a packet destined to output fibre j activates and deactivates the *request* signal identically as in I-PDBM. The *allow* signal is sent to every output module of the remaining output fibres (to every fibre if there is no packet) and to the output module of output fibre j associated with a delay $t > p$, where p is the shortest granted delay in previous iterations.

Step 2. Grant: New packets cannot leave the switch before precedent ones. Therefore, each output module (j, t) disables the signals from input modules associated with input fibre i if $\overline{z_{jt}}(i) = 1$. To guarantee the order of the packets that arrive within the same time slot, each output module (j, t) disables the signals from input module (i) if any input module with a round-robin precedent wavelength of the same input fibre has not activated *request* or *allow* signals for delay t. First, output modules keep activated the grants to the input modules that hold the *request* signal and deactivate the grants

DOI: 10.1002/ett

to the input modules that activated the *allow* signal. The number of available grants in the second step equals the number of available wavelengths in the fibre ($n-\overline{y_{jt}}$) minus the number of activated grants. Second, each (j, t) output module scans the enabled *request* signals and grants a delay to the first enabled ones that have not received a grant so far, according to grant pointer FG(j, t) and round-robin pointer WG(f) as in I-PDBM.

3.2. Algorithm justification

OI-PDBM converges when the signals become stabilised, i.e. there are neither new packet allocations nor assignments of better delays to granted packets. Like I-PDBM, OI-PDBM converges to a maximal size matching in min(M,nN) iterations at most.

Proof: (a) an output port only changes a *grant* signal if a previous input port *request* signal (according to the grant pointers) switches to an *allow* signal. A *request* signal only switches to an *allow* signal if the input module received a grant in the previous iteration from an output port that is associated with a shorter delay. Since the grants from delay 0 do not change after the first iteration, the algorithm converges in M iterations at most. (b) An input port is granted a shorter delay only if another input port was granted a shorter delay in the previous iteration. Since there are nN input ports, the algorithm converges in nN iterations at most.

OI-PDBM does not disorder packets. The output modules disable the signals from input modules if $\overline{z_{jt}}(i) = 1$, to respect order regarding scheduled packets of the same input–output pair. Also, an input module can receive a grant from an output module associated with delay t only if the precedent input modules of the same input fibre have been granted a delay (a) by the same output module (they have activated the *request* signal), (b) by another output module with a shorter delay (they have activated the *allow* signal) or (c) do not have a packet for the output fibre (they have also activated the *allow* signal). Therefore, the scheduler does not disorder the packets that arrive in the same time slot.

In this paper, we consider the round-robin packet order criterion. Note that the scheduler can be easily adapted to fulfil other criteria (such as the WASPNET criterion [8]).

3.3. Hardware implementation

Iterative parallel maximal size matching algorithms whose complexity is similar to that of I-PDBM have a decision time of a few nanoseconds for 256×256 ports [16, 17]. The changes in OI-PDBM to prevent mis-sequencing can be performed in a decentralised and parallel fashion, so they increase decision time, but do not directly affect scalability. Since the decision time for this architecture is $\sim$1 μs, hardware speed requirements are not a serious limitation.

The changes in OI-PDBM input modules regarding I-PDBM to activate the allow signal are simple and easy to implement. The new *allow* signal requires duplicating input–output module connections, which complicates the hardware but does not directly affect decision time.

The changes regarding the NM output modules are more complex. However, as we can see in Section 4, if n is large and N is small as in realistic backbone WDM topologies, M takes values of 2 or 3 and the number of required output modules is about 8 or 12. Each additional state vector $\overline{z_{jt}}(i)$ is handled in every time slot by just one output module (j, t) at a time. Moreover, if $M = 2$, the packets cannot leave the switch before previously arrived ones and $\overline{z_{jt}}(i)$ state vectors are not needed at all. If $M = 3$, we only need a unidirectional link between $j = 2$ and $j = 0$ output modules to signal if there were packets from each input fibre assigned in the previous time slot. For each input fibre and output module pair, the deactivation of input signals from the first port that do not activate any signal can be performed in parallel with a bus of length n carrying the logic NOR of the *allow* and *request* signals from the n input modules of the input fibre i and one comparator per input port with the round-robin pointer WG(i) suitably rotated. A simple hardware description for output module (j, t) is given in Figure 5. WG(i) is cyclically rotated at each comparator input in such a way that the bit corresponding to λ_k is the least significant one. If WG(i) = l in time slot d and any input port associated with a higher priority wavelength than λ_k in d does activate neither request nor allow signals, there will be at least a 1 in the A bits of the comparator input from l to k. Then, the comparator output will be 0 and the *request'* signal will be disabled. If the given input port received a grant in the previous iteration, the *request*' signal will also be disabled. If $\overline{z_{jt}}(i) = 1$, all the *request'* signals of the input ports will be deactivated. The *request'* signals in Figure 5(a) are the input signals for an ordinary I-PDBM output module. The only internal change of the output modules regarding I-PDBM is the calculation of the number of available wavelengths. In this case, the $n - y_{jt}$ value given by output module (j, $t+1$) in the accept phase of the previous algorithm execution has to be updated in every iteration. The available wavelengths are the remaining $n - y_{jt}$ wavelengths minus the number of ports with activated grant AND request signals can be calculated by an adder and a subtractor in a single step as shown in Figure 5(b).

Note that the figure simply shows that all the modifications can be performed in parallel, presumably increasing

DOI: 10.1002/ett

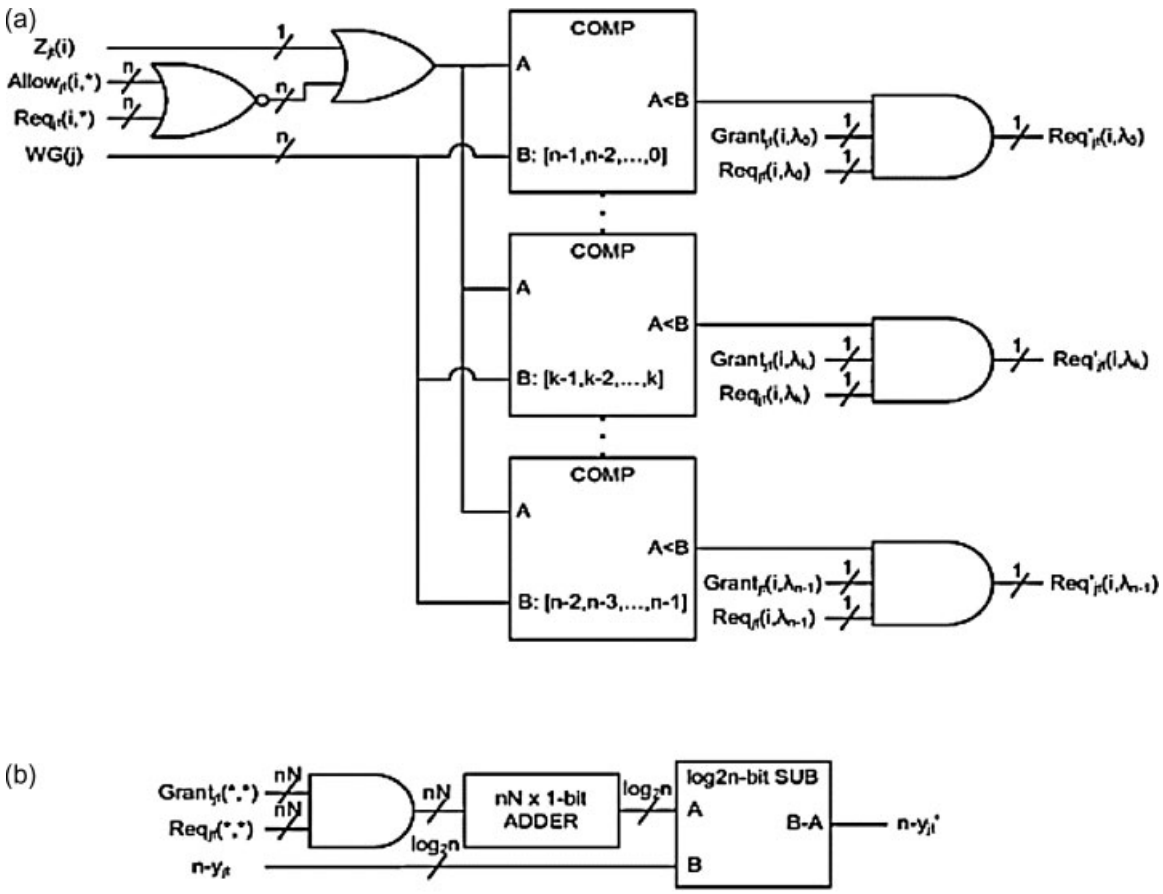

Fig. 5. Hardware changes of OI_PDBM output module (j, t) regarding I-PDBM: (a) input signal inhibitions and (b) calculation of available wavelengths.

the decision time but without sequential operations that limit scheduler scalability. In a real implementation, the combinational circuits in the figure can be further simplified. A real ASIC or FPGA implementation is beyond the scope of this paper.

4. RESULTS

In this section, we present simulation results to compare the proposed OI-PDBM algorithm with the previous I-PDBM algorithm and the OB architectures (see tables below) in terms of average delay, buffer requirements and practical convergence under uncorrelated (benign) or bursty traffic conditions. OB architectures are useful for comparison, as they provide the optimum throughput-delay performance, and do not disorder packets, by applying the scheduling algorithm in Reference [9]. The simulations were conducted using the oPASS tool [18].

The schedulers have been evaluated under n-SCWP traffic sources [14]. These sources consider an input fibre of n wavelengths as a batch traffic source of up to n-packets per time slot. As shown in Reference [14], n-SCWP sources are composed of the concatenation of a conventional discrete source (i.e. Bernouilli, bursty, etc.) working n times faster, and a round-robin dispatcher that spreads the (up to n) packets generated during each time slot across the fibre wavelengths.

Figures 6(a) and (b) show the average delay of OI-PDBM *versus* I-PDBM under n-SCWP Bernoulli traffic (OI-PDBM: solid line, I-PDBM: dotted line). Switch sizes are $N=\{2,4\}$, $n=\{2,8,32,64\}$, which corresponds to realistic backbone WDM topologies. Results obtained for higher values of $N=\{6,8\}$, not included in the paper,

DOI: 10.1002/ett

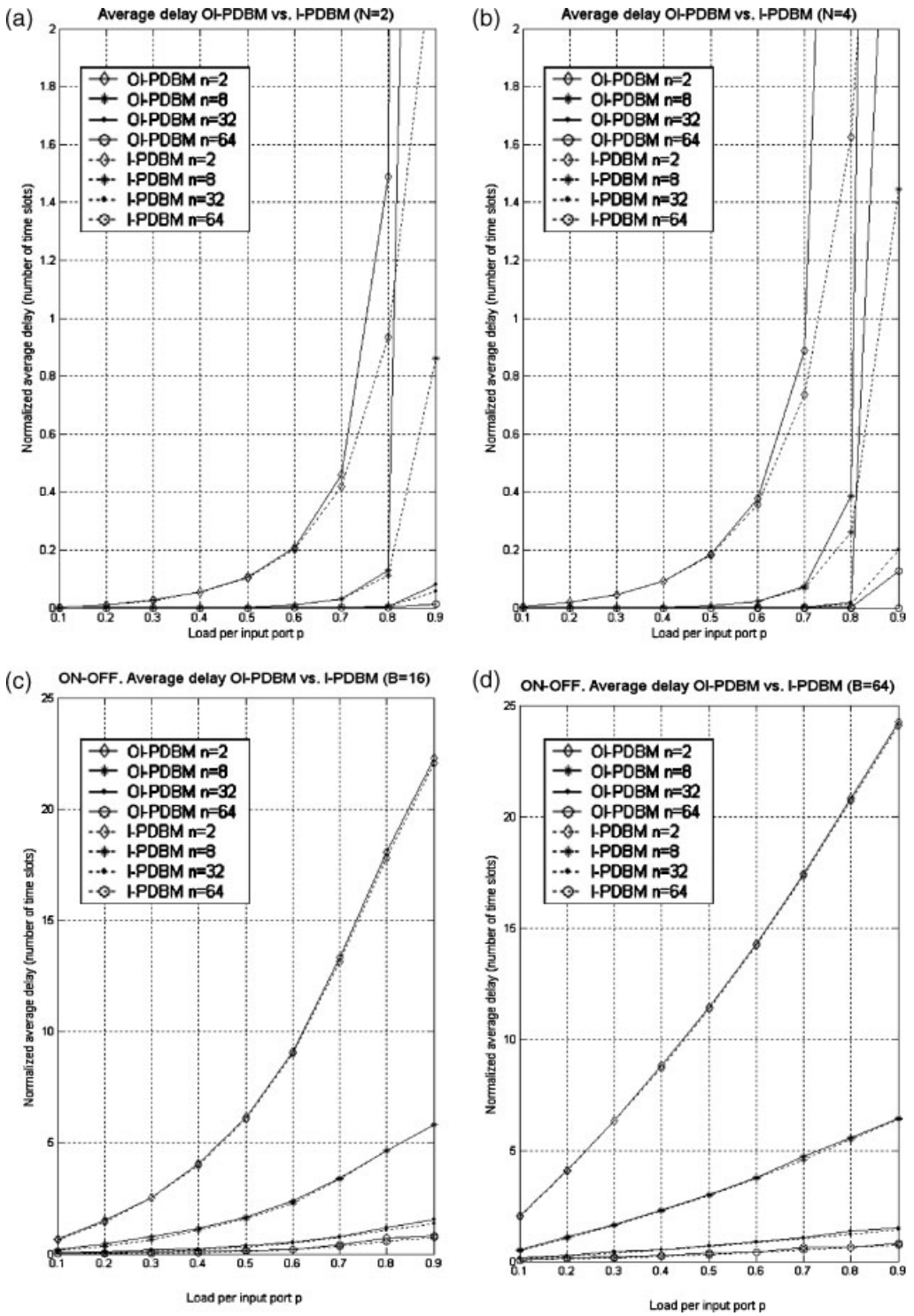

Fig. 6. (a, b) average delay under SCWP Bernoulli traffic; (c, d) average delay under SCWP MMPP traffic; (c) $\beta = 16$ and (d) $\beta = 64$ (OI-PDBM: solid line, I-PDBM: dotted line).

DOI: 10.1002/ett

Table 1. Buffer requirements (OB/I-PDBM/OI-PDBM). Bernoulli input traffic, 10^{-7} packet loss probability.

Switch size	$\rho=0.1$	$\rho=0.2$	$\rho=0.3$	$\rho=0.4$	$\rho=0.5$	$\rho=0.6$	$\rho=0.7$	$\rho=0.8$	$\rho=0.9$
$N=2, n=2$	2/2/2	3/3/3	3/3/3	4/4/4	5/5/5	5/6/6	7/8/8	10/10/14	18/20/30
$N=2, n=8$	1/1/1	2/2/2	2/2/2	2/2/2	2/2/2	2/2/2	3/3/3	3/4/4	6/8/9
$N=2, n=32$	1/1/1	1/1/1	1/1/1	1/1/1	2/2/2	2/2/2	2/2/2	2/2/2	2/3/3
$N=2, n=64$	1/1/1	1/1/1	1/1/1	1/1/1	1/1/1	1/1/1	2/2/2	2/2/2	2/2/2
$N=4, n=2$	3/3/3	3/4/4	4/4/4	5/5/5	6/6/6	7/8/8	9/11/11	14/16/20	26/30/30
$N=4, n=8$	1/1/1	2/2/2	2/2/2	2/2/2	2/2/2	3/3/3	3/3/3	4/5/6	8/10/13
$N=4, n=32$	1/1/1	1/1/1	1/1/1	1/1/1	2/2/2	2/2/2	2/2/2	2/2/2	3/3/5
$N=4, n=64$	1/1/1	1/1/1	1/1/1	1/1/1	1/1/1	2/2/2	2/2/2	2/2/2	2/2/3

do not differ from those shown and yield the same conclusions. Buffer sizes were adequate for OB architectures (packet loss probability below 10^{-9} under 90 per cent load): $M=\{35,10,3,2\}$ for $n=\{2,8,32,64\}$, respectively. We can observe similar performance in the average delay of OI-PDBM and I-PDBM for low and medium input loads. For very high input loads and low n, I-PDBM is better than OI-PDBM since delays longer than the shortest ones with no input port and output fibre contention have to be selected to maintain the packet sequence. To illustrate the effect of traffic burstiness, figures 6(c) and (d) depict the average delay of OI-PDBM and I-PDBM (OI-PDBM: solid line, I-PDBM: dotted line) under a n-SCWP arrival Markov-modulated ON-OFF Poisson process (MMPP), for burst lengths of $\beta=16$ (Figure 6(c)) and $\beta=64$ (Figure 6(d)). Switch sizes are $N=4$, $n=\{2,8,32,64\}$, and buffer sizes are the same as above. Bursty traffic affects OI-PDBM performance as in the case of I-PDBM and OB architectures [14]. Therefore, except for Bernoulli traffic at high loads and low number of wavelengths (unrealistic scenarios), the average delay of OI-PDBM is very similar to that of I-PDBM.

Table 1 shows buffer requirements for a packet loss probability of 10^{-7} under Bernoulli traffic for OB, I-PDBM and OI-PDBM (simulation length is 10^9 packets). This is a good feasibility metric for OPS nodes because FDL length is a serious bottleneck nowadays. We observe that OI-PDBM buffer lengths are very small as in the I-PDBM case and the ideal OB scenario, except for $n=2$ and high loads. For these scenarios the order constraints become stronger and packet assignments to higher delays than necessary to maintain packet order increase buffer requirements.

Tables 2 and 3 compare the theoretical convergence bound ($\min(nN, M)$) with the number of iterations K to converge with a probability of above $1–10^{-6}$ (90 per cent input load). I-PDBM and OI-PDBM behave almost identically, and in all cases the number of iterations is quite low.

Table 2. Practical number of iterations for I-PDBM/OI-PDBM convergence *versus* theoretical convergence bound (bold), Bernoulli traffic.

Bernoulli $\rho=0.9$	$n=2$	$n=8$	$n=32$	$n=64$
$N=2$	3/3 **4**	3/3 **10**	2/2 **3**	2/2 **2**
$N=4$	5/4 **8**	3/3 **10**	2/3 **3**	2/2 **2**

Table 3. Practical number of iterations for I-PDBM/OI-PDBM convergence *versus* theoretical convergence bound (bold), MMPP traffic.

MMPP $\rho=0.9, N=4$	$n=2$	$n=8$	$n=32$	$n=64$
$\beta=16$	5/5 **8**	6/6 **10**	3/3 **3**	2/2 **2**
$\beta=64$	5/5 **8**	6/6 **10**	3/3 **3**	2/2 **2**

5. CONCLUSIONS

In this paper, we propose the OI-PDBM parallel iterative matching scheduler for IBWR optical packet switches. OI-PDBM is simple and allows practical parallel hardware implementation. It performs well in terms of delay, buffer requirements and convergence speed and guarantees packet order. Therefore, this scheduler efficiently eliminates the packet mis-ordering in IBWR switches, and endorses the application of this architecture in OPS networks as a feasible alternative to less scalable OB architectures.

ACKNOWLEDGEMENTS

This research has been supported by project grants CON-PARTE TEC2007-67966-C03-01/02/TCM (MEC, Spain) and MIND-GAP-5 PGIDIT08TIC010CT (Xunta de Galicia, Spain). The work described in this paper was carried out with the support of the BONE-project ('Building the Future Optical Network in Europe'), a Network of Excellence funded by the European Commission through the 7th ICT-Framework Programme. The authors

participate in the TEC2008-02552-E thematic network (MEC, Spain).

REFERENCES

1. Dittman L, Develder C, Chiaroni D, *et al*. The European IST project DAVID: a viable approach toward optical packet switching. *IEEE Journal of Selected Areas in Communications* 2003; **21**(7): 1026–1040.
2. Hunter DK, Nizam MHM, Chia MC, *et al*. WASPNET: a Wavelength switched packet network. *IEEE Communications Magazine* 1999; **37**(3): 120–129.
3. Pavon-Mariño P, García-Haro J, Malgosa-Sanahuja J, Cerdán F. Scattered versus shared wavelength path operation, application to output buffered optical packet switches. A Comparative Study, *SPIE/Kluwer Optical Networks Magazine* 2003; **4**(6): 134–145.
4. Pavón-Mariño P, García-Haro J, Malgosa-Sanahuja J, Cerdán F. Maximal matching characterization of optical packet input-buffered wavelength routed switches. In *Proc. of 2003 IEEE Workshop on High Performance Switching and Routing (HPSR 2003)* Torino, Italy, June 2003; pp. 55–60.
5. Bennett JCR, Partridge C, Shectman N. Packet reordering is not pathological network behavior. *IEEE/ACM Transactions on Networking* 1999; **7**(6): 789–798.
6. Blanton E, Allman M. On making TCP more robust to packet reordering. *ACM Computer Communication Review* 2002; **32**(1): 20–30.
7. Callegati F, Cerroni W, Raffaelli C. Impact of optical packet loss and Reordering on TCP performance. In *Proc. of IEEE Global Telecommunications Conference (Globecom)*, San Francisco, CA, November 2006.
8. Chia MC, Hunter DK, Andonovic I, *et al*. Packet loss and delay performance of feedback and feed-forward arrayed-waveguide gratings-based optical packet switches with WDM inputs-outputs. *IEEE Journal of Lightwave Technology* 2001; **19**(9): 1241–1254.
9. Pavón-Mariño P, González-Castaño FJ, García-Haro J. Round-robin wavelength assignment: a new packet sequence criterion in optical packet switching SCWP networks. *European Transactions on Telecommunications* 2006; **17**(4): 451–459.
10. Zhong WD, Tucker RS. Wavelength routing-based photonic packet buffers and their applications in photonic packet switching systems. *IEEE Journal of Lightwave Technology* 1998; **16**(10): 1737–1745.
11. Takahashi H, Suzuki S, Kato K, Nishi I. Arrayed-waveguide grating for wavelength division multi/demultiplexer with nanometre resolution. *Electronics Letters* 1990; **26**: 87–88.
12. Guillemot C, Renaud M, Gambini P, *et al*. Transparent optical packet switching: the European ACTS KEOPS project approach. *IEEE Journal of Lightwave Technology* 1998; **16**(12): 2117–2134.
13. Rodelgo-Lacruz M, Pavón-Mariño P, González-Castaño FJ, García-Haro J, López-Bravo C, Veiga-Gontán J. Enhanced parallel iterative schedulers for IBWR optical packet switches. In *Proc. of 11th Conference on Optical Network Design and Modelling (ONDM)*. May 2007; Athens (Greece).
14. Pavón-Mariño P, García-Haro J, Jajszczyk A. Parallel desynchronized block matching: a feasible scheduling algorithm for the input-buffered wavelength-routed switch. *Computer Networks* 2007; **51**(15): 4270–4283.
15. McKeown N. The iSLIP scheduling algorithm for input-queued switches. *IEEE/ACM Transactions on Networking* 1999; **7** (2): 188–201.
16. Chao HJ. Saturn: a terabit packet switch using dual round-robin. *IEEE Communication Magazine* 2000; **38**(12): 78–84.
17. Hung CK, Hamdi M, Tsui C. Design and implementation of high-speed arbiter for large scale VOQ crossbar switches. *Proc. of Int. Symposium on Circuits and Systems (ISCAS)* vol. 2, 2003; pp. 308–311.
18. Bueno-Delgado MV, Veiga-Gontán JA, Pavón-Mariño P, García-Haro J. oPASS: a simulation tool for the performance evaluation of optical packet switching architectures. In *Proc. of European Symposium on Simulation Tools for Research and Education in Optical Networks (STREON)* October 2005, Brest (France).

AUTHORS' BIOGRAPHIES

Miguel Rodelgo Lacruz received his Telecommunications Engineering degree in 2004 and his Advanced Studies Diploma in 2006 both from the University of Vigo, Spain. He has worked as system administrator with Comunitel Global, Spain, as an invited professor with University of Vigo and as R&D Engineer with the GTI group of the same university, where he held a FPI grant. He is currently a project manager with Gradiant, Spain. He is interested in high performing networks, switching systems and performance analysis. He has published several papers in conferences and international journals and has collaborated in diverse national and international projects.

Pablo Pavon-Mariño received the Telecommunication Engineering degree in telecommunications in 1999 from the University of Vigo (UVIGO), Spain. In 2000, he joined the Technical University of Cartagena (UPCT), where he is an Associate Professor with the Department of Information Technologies and Communications. He received the Ph.D. degree from this University in 2004. His research interests include performance evaluation, planning and optimization of communication networks.

Francisco J. González-Castaño received the Ingeniero de Telecomunicación degree from the University of Santiago de Compostela, Spain, in 1990 and the Doctor Ingeniero de Telecomunicación degree from the University of Vigo, Spain, in 1998. He is currently a Catedrático de Universidad (Full Professor) with the Departament of Telematics Engineering, University of Vigo, where he leads the Information Technologies Group (http://www-gti.det.uvigo.es). He is also currently with Gradiant, Spain, as the Network Research Director. He has authored more than 50 papers published in international journals, in the fields of telecommunications and computer science, and has participated in several relevant national and international projects. He is the holder of three Spanish patents, one European patent, and one U.S. patent.

DOI: 10.1002/ett

Joan García-Haro is a Professor at the Polytechnic University of Cartagena, Spain. He is author or co-author of more than 60 journal papers mainly in the fields of switching, wireless networking and performance evaluation. From April 2002 to December 2004 he served as EIC of the IEEE Global Communications Newsletter, included in the IEEE Communications Magazine. He is Technical Editor of the same magazine from March 2001. He also holds an Honorable Mention for the IEEE Communications Society Best Tutorial paper Award (1995).

Cristina López-Bravo received her PhD degree in Telecommunications Engineering from the Polytechnic University of Cartagena (Spain) in 2004. Since October 2005 she has been an Assistant Professor with the Telematics Engineering Department at University of Vigo (Spain), and a researcher with the GTI Group of the same university. Her research interests include Optical Packet Switching and High-Performance Networking with focus on design, development and evaluation of packet scheduling algorithms.

Juan Antonio Veiga-Gontán received the Telecommunication Engineering degree in 2001 from the University of Vigo, Spain. Since 2005, he holds a FPI grant at the Department of Information Technologies and Communications, Polytechnic University of Cartagena (UPCT), Spain. His research focuses on optical networks.

Felipe Gil-Castiñeira received the M.Sc. degree in telecommunication engineering (major in telematics) and the Ph.D. degree in telecommunication engineering from the University of Vigo, Vigo, Spain, in 2002 and 2007, respectively. He is currently an Assistant Professor with the Department of Telematics Engineering, University of Vigo. His research interests include wireless and car-to-car communication technologies, embedded systems, nomadic devices, wireless sensor networks, and ubiquitous computing.

Carla Raffaelli received the M.S. and the Ph.D degrees in electrical engineering and computer science from the University of Bologna, Italy, in 1985 and 1990, respectively. She is associate professor in switching systems and telecommunication networks at the University of Bologna. She has been with the Department of Electronics, Computer Science and Systems of the University of Bologna, Italy, since 1985 where she became a Research Associate in 1990. Her research interests include performance analysis of telecommunication networks, switching architectures, protocols and broadband communication. Since 1993 she participated in European funded projects on optical packet-switched networks, the RACE- ATMOS, the ACTS-KEOPS and the IST-DAVID projects. She is now active in the EU funded e-photon/One network of excellence. She also participated in many national research projects on telecommunication networks. Prof. Raffaelli is the author of many technical papers on broadband switching and network modelling and regularly acts as a reviewer for top international conferences and journals. She is author or co-author of more than 100 conference and journal papers mainly in the field of optical networking and performance evaluation.

DOI: 10.1002/ett