Scheduling variable-size packets in the DAVID metroplitan area network

Scheduling Variable-Size Packets in the DAVID Metropolitan Area Network

A. Bianco, J.M. Finochietto, F. Neri, and V. Sarra
Politecnico di Torino
Corso Duca degli Abruzzi 24, 10129 Torino, Italy
Email: {bianco, finochietto, neri}@polito.it

G. Galante
Istituto Superiore Mario Boella
via Pier Carlo Boggio 61, 10138 Torino, Italy
Email: galante@ismb.it

***Abstract*— DAVID is a research project sponsored by the European Union aimed at the design of an optical packet-switched network for the transport of IP traffic. The DAVID network has a two-level hierarchical structure, with a backbone of optical packet routers inter-connected in a mesh, and metropolitan areas served by sets of optical rings interconnected by passive memoryless devices called Hubs.**

The paper focuses on the metropolitan area network and its components: the nodes and the Hub. Access is regulated by a dynamic time-division multiple-access scheme allocating slots in sets of wavelengths that provide multi-channel pipes among ring pairs. This paper proposes a new resource allocation scheme capable of transporting variable-size packets without segmentation into fixed-size data units. Resource sharing among nodes is granted by two scheduling algorithms running on different time scales: the first one is centralized at the Hub and provides coarse connectivity among ring pairs; the second one runs at nodes and provides finer node-to-node connectivity.

The Hub scheduling algorithm is derived from well known algorithms in the literature; a novel heuristic scheduling algorithm running at nodes is proposed for datagram (not-guaranteed) traffic and its performance is studied by simulation.

I. Introduction

The DAVID (Data And Voice Integration over Dense-wavelength division multiplexing) project is part of the Information Society Technology (IST) Program sponsored by the European Union. Its aim is the design of an optical packet-switched network for the transport of IP traffic over metropolitan, national and international distances.

The major issues addressed by DAVID are:

- the design of an optical network offering a transport format independent of the traffic type: the clients of the DAVID network are mainly IP routers and/or switches that collect traffic from legacy networks;
- the evaluation of the best areas of applicability of optics and electronics in order to find the optimum mix of technologies for future very-high-capacity networks;
- the careful definition of an evolution strategy so as to ensure a smooth transition from the current to the future network infrastructure.

The DAVID network [1] has a two-level hierarchical architecture where several optical-ring metropolitan area networks (Metros) are interconnected by a Wide Area Network (WAN) backbone.

The network operates in optical packet-switched mode and most of the project is focused on a synchronous network operation and control, and on the transport of fixed-size packets, because this greatly simplifies high-speed operations in both the optical and the electronic domains. Thus, variable-size IP packets must be fragmented by the sender before being transmitted on the fiber, and reassembled at the receiver before being delivered to the intended recipient. In [2] and [3] we presented media access control (MAC) protocols for transporting on the DAVID Metro best-effort and guaranteed traffic fragmented into fixed-size packets; in this paper we explore instead solutions for accommodating variable-size packets in the DAVID optical metro network.

The rest of the paper is organized as follows. In Section II we briefly review the general architecture of the DAVID network. In Section III we focus on the Metro describing in more detail its operations, and propose our new resource allocation scheme. In Section IV we present selected simulation results to assess the performance of the proposed scheme. We conclude the paper in Section V, where we give future research directions.

II. General Overview

As shown in Fig. 1, the DAVID backbone network consists of optical packet routers interconnected by a mesh network, while each Metro network comprises one or more rings interconnected through a bufferless Hub.

Each ring collects traffic from several nodes; each Hub interconnects a number of rings, and is connected to an optical packet router in the WAN through a gateway. Access points to the network are provided both in DAVID Metro

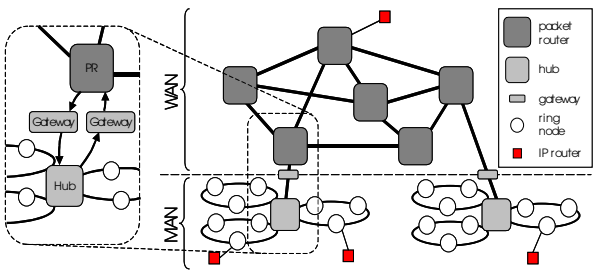

Fig. 1. General overview of the DAVID network.

and backbone, and the traffic is collected by IP routers and switches connected to local area networks.

Bandwidth partitioning in the Metro is obtained by a superposition of a wavelength division multiplexing (WDM) and a time division multiplexing (TDM) access protocol: each fiber carries several data wavelength channels at 10 Gbit/s which are shared in the time domain to provide connectivity among rings.

In packet-switched networks buffering inside routers is needed to solve contentions arising among packets arriving in a given node and headed to the same output port. In the DAVID WAN, optical packet routers provide buffering in the optical domain by means of fiber delay lines. No packet buffering in the optical domain is instead performed for packets flowing among ring nodes in the same Metro network. Indeed, packets are buffered in ring nodes in the electrical domain, and are sent on the Metro network only when there are enough free resources on the Metro to travel from source to destination without being stored at any intermediate node. Thus, buffers are pushed towards the edge of the Metro network and sharing of ring resources among nodes must be regulated by a properly designed MAC protocol.

III. Resource Allocation in the Metro Network

The Metro network has been designed to provide a seamless transition from the present situation, where telephone and data networks are heavily dependent on semi-static TDM-based synchronous digital hierarchy (SDH)/synchronous optical network (SONET) ring architectures, to a future scenario, where more dynamical, optical packet-switched network will take the lion's share.

A. WDM access scheme

Each fiber conveys W (typically equal to 32) wavelengths at 10 Gbit/s called *data channels,* plus wavelengths at 2.5 Gbit/s for signalling purposes dubbed *signalling* or *control channels*. The separation induced by such an out-of-band signalling scheme allows to keep bulk data in the optical domain as long as possible and to process them electronically only when they are received. The price that has to be paid is that the control channel must be terminated, electronically processed and retransmitted at each node, and this justifies its lower bitrate.

The Metro consists of several uni-directional optical rings interconnected by a bufferless Hub. Protection and restoration may be deployed at the price of increased node complexity by adding to each ring a counter-rotating path on which data are switched by nodes in case of failure.

Each node in the Metro is connected to a *physical ring* formed by one or more fibers running in parallel; yet, in general, a node can only access a subset of contiguous wavelengths selected among these circulating on a fiber called *waveband*. This allows to partition a physical ring into several *logical rings* containing only some of the nodes connected to the same physical ring. Each logical ring comprises a control channel serving its data channels, and we assume that nodes connected to different logical rings cannot forward packets directly among them; therefore, packets are routed among different logical rings at the Hub.

The number of data channels in a logical ring is typically limited to 4 or 8, because nodes are equipped with integrated optical modules with limited tuneability comprising one tunable *data transmitter* and one tunable *data receiver*. In addition, to simplify the access to the control channel, each node needs a dedicated *signalling transceiver* which is always tuned on the control channel independently of the data transceivers behavior.

The DAVID metro operates in a synchronous, time-slotted fashion. Time slots, whose typical duration is 1 μs, are the minimum granularity in resource allocation, and typically rule the dynamics of network control, in the sense that both node access decisions and Hub switching happen between slot boundaries. In this paper we propose instead to decouple the temporal dynamics of Hub switching and node access: Hub ring-to-ring interconnection patterns are held for several time slots, in which nodes can fit variable-size packets treated as trains of fixed-size data units. This approach has a cost in terms of delays and efficiency in network utilization, but presents advantages because processing of variable-size packets is simplified, and technological constraints at the Hub are relaxed.

B. Hub Scheduling: Coarse TDM Scheme

The time dimension is used to set up at the Hub ring-to-ring bandwidth pipes among different logical rings using a TDM access scheme. Hub operations are based upon fixed-size *Hub-slots* lasting T_{Hub}. Hub-slots are further partitioned in a number of *node-slots,* as explained later.

In DAVID demonstrators, the Hub is implemented with a broadcast-and-select structure [4] using Semiconductor Optical Amplifiers (SOAs) as a switching technology. We assume here that the Hub is reconfigured only on Hub-slot boundaries, relaxing technological constraints, and permitting other, more consolidated switching technologies.

Since the Hub is a bufferless all-optical TDM switch, switching is performed with a waveband granularity, implementing *permutations* between logical rings: no two input (or output) rings can be connected to the same output (or input) ring in a given Hub-slot. Thus, all packets traveling on an input logical ring (i.e., on a waveband) are transparently transferred to the same destination logical ring. This operation assumes that all wavebands comprise the same number of wavelengths, and we make this assumption in the sequel of the paper.

According to the assumptions above, the Hub switching granularity is therefore coarse both in the time (long Hub-slots) and in the wavelength (waveband to waveband switching) domains.

A scheduling algorithm drives the Hub through a sequence of input/output configurations. To make the scheduling algorithm simpler, we assume without loss of generality that the propagation delay along all the physical (and therefore logical) rings is equivalent to the same integer number of Hub-slots. This constraint can be removed either making all the rings

the same length by adding fiber delay lines, or by taking into account the propagation delay on each ring, and delaying accordingly data at the Hub.

A sequence of F_{Hub} Hub-slots is organized into a fixed-length frame, where F_{Hub} is designed according to the desired ring-to-ring connection granularity. The scheduling for a given ring-to-ring traffic matrix can be computed at the Hub by using standard techniques based on iterated applications of modifications of the maximum size/maximum weight matching algorithms (see for instance, [5], [6]). For simplicity, in this paper we assume that the ring-to-ring traffic matrix is known in advance; if this is not the case, it can be estimated with a measurement algorithm, as discussed in [2]. Note that this assumption scales well with the network size, since its complexity depends on the number of rings (and not of nodes) in the Metro.

The Hub-slot allocation is distributed to nodes on the control channels as follows: the first few bits transmitted on each control channel at the beginning of a new Hub-slot denote the destination ring to which the data in the corresponding waveband will be forwarded upon reaching the Hub.

C. Node Scheduling: Fine TDM Scheme

A second level of time multiplexing is implemented at nodes: a node's allocation granularity is typically finer than the Hub's since the node architecture is simpler and it can be more agile. All nodes operate synchronously with the time reference established by the transit of Hub-slots: each Hub-slot is further divided into an integer number $F_{\mathrm{node}} = T_{\mathrm{Hub}}/T_{\mathrm{node}}$ of shorter *node-slots* lasting T_{node} each. Node transceivers can tune very quickly (i.e., in a time $\ll T_{\mathrm{node}}$) on any wavelength in the waveband(s) to which the node is attached, and tuning can happen only on node-slot boundaries.

Several alternatives were proposed in the DAVID project for the Metro node architecture. We consider here the scheme described in [7], in which nodes are attached to data channels on the rings with simple passive couplers. This makes the node much cheaper and permits to cascade a larger number of nodes (easily more than 16 nodes) on a typical ring. A single-channel receiver and a single-channel transmitter can be tuned to any wavelength in the waveband, in order to receive and transmit at most one packet in a given node time slot.

Since packets cannot be erased from the channel, two separate wavebands are used: one for transmission and one for reception, requiring to double the number of wavelengths on the rings (without however requiring any increase in the amount of information that is switched at the Hub and at nodes). Switching at the Hub must shift received information from the transmission waveband of the input ring to the reception waveband of the reception ring.

Packets traveling on data channels are delayed with a fiber delay line so as to allow the signalling receiver to terminate the control channel and extract state information for the current Hub-slot, to give the node's CPU enough time to schedule packet transmissions in the current Hub-slot, and to regenerate the updated information in the signalling channel on the outgoing fiber.

In this framework, node-slots transmitted on the signalling channel at the beginning of each Hub-slot must describe the time/wavelength allocation map of all node-slots transported on the data channels in the same Hub-slot. Based upon this rich information on the state of each Hub-slot, variable-length packets can be transmitted without segmentation, treating them as trains of slots: each packet occupies a run of contiguous node-slots, possibly transmitted on different wavelengths in the same Hub-slot. However, since at any given time each node can receive data on at most one wavelength in the same waveband, source nodes must make sure that there is no time overlapping among packets sent to the same destination node, by refraining from transmitting whenever this may happen.

The proposed node access protocol acts as follows. A packet consisting of l node-slots can be sent on data channels only when both the following conditions hold:

1) it is possible to find on the data channel in the current Hub-slot a sequence of l contiguous node-slots, possibly on different wavelengths;
2) none of the previously found node-slots already contains, on a different wavelength, any (portion of a) packet addressed to the considered destination node.

It is easy to observe that the Metro behaves as a distributed input-queued packet switch where buffers are located at node data transmitters; it thus suffers from the well-known head-of-line (HOL) blocking phenomenon [8], which, however, can be easily overcome by providing virtual output queues (VOQ) in the electronic interfaces driving the data transmitters. Note that the HOL blocking can be completely removed from the Metro only when adopting a per-destination-node VOQ policy; yet, we preferred to opt for a per-destination-ring VOQ policy to improve network scalability. Simulation results show that the performance degradation for using such non-optimal queueing architecture is minimal.

The node scheduling algorithm must be as simple as possible since it must be executed in a few hundred nanoseconds. Hence, after reading from the control channel the destination ring to which the Hub-slot will be forwarded, the algorithm checks the destination node to which the packet at the head of the corresponding VOQ is addressed and scans from the beginning the slot allocation map looking for a "hole" satisfying conditions 1 and 2 above. If it is found, the packet is allocated and the scan continues from the current position looking for a "hole" for the next packet in the same VOQ. Such procedure is repeated until either the VOQ is empty or the end of the Hub-slot is reached.

This scheduling can be seen as a generalization to the case $T_{\mathrm{Hub}} > T_{\mathrm{node}}$ of the MAC protocol proposed in [2]. In [3], the computation of the two TDM schedules is centralized at the Hub and support for two classes of traffic is provided.

D. Throughput Fairness

The node scheduling algorithm presented above can exhibit fairness problems under unbalanced traffic; this is due to the

ring topology, in which upstream nodes have generally better access chances than downstream nodes.

Credit-based schemes, such as the Multi-MetaRing [9], previously studied in the context of a single ring, can enforce throughput fairness. MetaRing [10] was proposed by Y. Ofek for ring-based electronic networks: see [2], [11] for details on how it is implemented in the DAVID Metro.

IV. Simulation Results

A. Network Parameters

The simulated network consists of 4 rings each connected to 15 nodes; thus the Hub has 4 ports.

Each ring conveys $8 + 1$ wavelengths: 4 for transmission, 4 for reception and 1 for signaling purposes. Each node is equipped with 1 tunable data transceiver for accessing the $4+4$ data channels and 1 fixed signalling transceiver for accessing the control channel. Each node's VOQ can store up to 100 000 packets, and nodes can access all wavelengths with granularity $T_{\mathrm{node}} = 100$ ns. Instead, the Hub switching granularity T_{Hub} can be varied between 0.1 μs and 12.8 μs; thus, one Hub-slot contains between 1 and 128 node-slots. The Hub frame length is set to $F_{\mathrm{Hub}} = 400$. Each ring contains 8192 node-slots corresponding to ≈ 160 km, and the distance between the nodes has been fixed to 512 node-slots (≈ 10 km).

B. Traffic Model

The traffic on the Metro is described according to a 4×4 matrix $\mathbf{R}$, representing the probability that a packet generated by a node attached to ring i is sent to ring o. All source nodes generate the same amount of traffic, measured in terms of number of node-slots; packet inter arrival times are geometrically distributed and the packet length is uniformly distributed between 1 and 8 node-slots. Packets addressed to ring o are uniformly distributed among all nodes connected to ring o.

We consider two traffic scenarios, named *uniform* and *unbalanced*. In the *uniform* traffic pattern $\mathbf{R}_{i,o} = 1/4 \ \forall i, o$. For the *unbalanced* traffic pattern $\mathbf{R}_{i,o} = 0.7$ when $i = o$, and $\mathbf{R}_{i,o} = 0.1$ otherwise; in other words, the ratio among intra-ring traffic and inter-ring traffic is 7. For simplicity, we assume that the Hub scheduling is pre-computed and matched to the traffic pattern.

C. Numerical Results

We analyze by simulation throughput and queueing delays under different traffic patterns, as a function of the offered load. Throughput (and offered load) are normalized with respect to the network upstream capacity, and throughput is measured as the ratio between used and available node-slots. Queueing delays are measured in node-slots, starting from when the packet is generated and inserted in the proper queue until the last bit of the packet is transmitted on the ring.

All the presented curves show steady-state values computed from statistically significant measures obtained by simulation. We plot graphs for a single source ring, namely ring 1, but the same behavior holds for all other rings due to traffic symmetry. Nodes on the same ring do not exhibit throughput unfairness thanks to the MetaRing algorithm.

In Fig. 2 we plot the total throughput on ring 1 for variable values of F_{node}. Throughput increases linearly from 0 to 0.8 as the offered load on ring 1 varies from 0 to 0.8, then saturates to values around 0.8. The inability to fully exploit the available bandwidth is due to two factors: first, the sub-optimal node scheduling algorithm does not allow nodes to fully exploit all available node-slots due to its simplicity; second, the implemented version of the Metaring algorithm, needed to induce throughput fairness among nodes on the same ring, does not allow nodes to fully exploit network resources due to the single transceiver choice made in node architecture. Improvements of the considered fairness control scheme are outside the scope of this paper.

The curve for $F_{\mathrm{node}} = 1$ refers to the case in which variable-size packets are transmitted in interleaved segments, without enforcing contiguity; as expected, the throughput achieved in these conditions is the highest. When packet contiguity is enforced, it can be observed that the maximum throughput increases by increasing the value of F_{node}. This is due to a quantization effect: since the algorithm requires to find a set of contiguous node-slots to allocate a packet transmission, filling the last portion of the Hub-slot becomes difficult; this effect has a more significant impact on performance for shorter Hub-slot sizes.

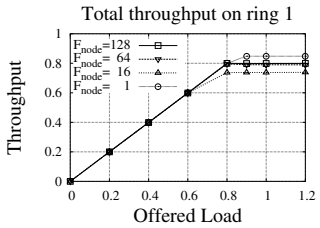

Fig. 2. Total throughput on ring 1 under uniform traffic for different values of F_{node}.

In Fig. 3 we plot total throughput (left) and per destination ring throughput (right) on ring 1 under unbalanced traffic pattern for different values of F_{node}. Observations similar to those outlined for the uniform traffic pattern hold.

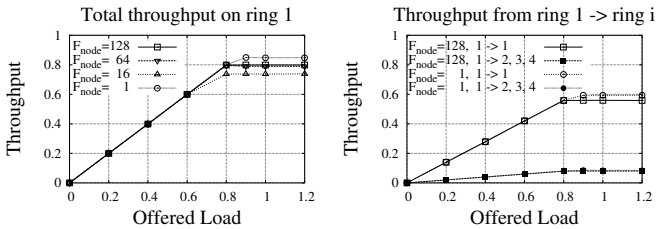

Fig. 3. Total throughput (left) and per destination ring throughput (right) on ring 1 under unbalanced traffic for different values of F_{node}.

Finally, in Fig. 4, we plot the queueing delays on ring 1 toward ring 1 (left) and ring i (right) for unbalanced traffic pattern, for $F_{\mathrm{node}} = \{1, 128\}$. As expected, increasing the Hub

switching time has an important effect on queueing delays, particularly at low loads; queueing delays increase by an amount of slots close to the number of node-slots comprised in a Hub-slot. Indeed, on average, a node has to wait a significantly longer time to access network resources when $F_{\rm node} = 128$ instead of when $F_{\rm node} = 1$, due to the coarser granularity in slot allocation at the Hub.

Note, however, that delays are less dependent on the offered load (i.e., curves are flatter) in the case of larger frames, as typical for the frame-based operation of switching systems (see for example [12] for another instance of this behavior).

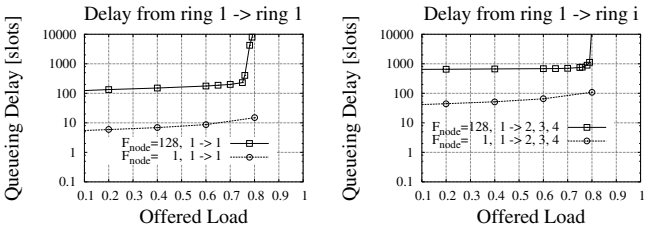

Fig. 4. Queueing delay from ring 1 to ring 1 (left) and from ring 1 to other rings (right) under unbalanced traffic for $F_{\rm node} = 1$ (circle marker) and $F_{\rm node} = 128$ (square marker).

V. Conclusions and Future Work

In this paper we have proposed network control and resource allocation algorithms for a Metro WDM network capable of transporting variable-size data units in the optical domain.

In addition to the proposed scheme for handling variable-size packets, an interesting contribution of the paper is the proposal of controlling Hub switching configurations on a different time scale with respect to access decisions taken by nodes along the WDM rings. In particular, in the considered Metro architecture we have:

- a complex and relatively slow Hub, which computes locally the scheduling of ring-to-ring permutations on windows lasting about 10 μs;
- simpler and more agile terminals, which compute in a distributed fashion the scheduling in each permutation window, using slots lasting around 100 ns.

This is equivalent to introducing in our optical packet-switching environment a decoupling between the control plane dynamics (at the Hub) and the access/transmission dynamics (at node interfaces), which is well in line with current trends of optical networking. The proposed approach is also motivated by the fact that network control dynamics are mostly related to quality of service (QoS) requirements, which do not vary with the never-ending increase of transmission data rates. The price to pay in terms of performance due to the introduction of this coarse Hub control granularity is a minor throughput degradation, and a more significant delay increase at low loads, as shown by simulation results. This is to be traded with higher technological requirements at the Hub, and the packet segmentation/reassembly complexity.

In the continuation of our work, in addition to studying the network behavior with different traffic scenarios, we will focus on the design of simple and efficient heuristic algorithms for locally scheduling variable-size packets at node interfaces, and on the design of fairness control schemes with better delay and throughput properties than Metaring extensions that were considered in this paper.

Acknowledgment

This work has been funded as part of the European Union IST DAVID project.

References

[1] C. Develder, M. Pickavet, N. Leligou, F. Callegati, F. Neri, "The European IST Project DAVID: a Viable Approach Towards Optical Packet Switching," *Journal on Selected Areas in Communications*, vol. 21, no. 7, Sept. 2003.

[2] A. Bianco, G. Galante, E. Leonardi, and F. Neri, "Measurement Based Resource Allocation for Interconnected WDM Rings," *Photonic Network Communications*, vol. 5, no. 1, pp. 5–22, Jan. 2003.

[3] A. Bianco, J. Finochietto, E. Leonardi, P. Mitton, F. Neri, L. Quarello, "Multiclass Resource Allocation in Interconnected WDM Rings," *7th Working Conference on Optical Network Design and Modeling (ONDM 2003)*, Febr. 2003, Budapest, Hungary.

[4] D. Chiaroni, et al., "First demonstration of an asynchronous optical packet switching matrix prototype for Multi-Terabit-Class Routers/Switches," *postdeadline paper at ECOC 2001*, Vol. 6, PD.A.1.8, pp. 60-61, 2001.

[5] B. Hajek, T. Weller, "Scheduling Non-Uniform Traffic in a Packet-Switching System with Small Propagation Delay," *IEEE Transactions on Networking*, Vol. 5, No. 6, Dec. 1997, pp. 813–823.

[6] C.S. Chang, W.J. Chen, H.Y. Huang, "Birkhoff-von Neumann Input Buffered Crossbar Switches," *IEEE Conference on Computer Communications (INFOCOM 2000)*, Tel Aviv, Israel, Mar. 2000, pp. 1614–1623.

[7] A. Stavdas, "Architectures, Technology and Strategies for a Gracefully Evolving Optical Packet Switching Networks," *Optical Networks Magazine*, Vol. 4, No. 3, May/June 2003.

[8] M. Karol, M. Hluchyj, S. Morgan, "Input Versus Output Queuing on a Space Division Switch," *IEEE Transactions on Communications*, Vol. 35, No. 12, Dec. 1987, pp. 1347–1356.

[9] M. Ajmone Marsan, A. Bianco, E. Leonardi, A. Morabito, F. Neri, "All-Optical WDM Multi-Rings with Differentiated QoS," *IEEE Communications Magazine, Feature topic on Optical Networks, Communication Systems and Devices, M. Atiquzzaman, M. Karim (eds.)*, Vol. 37, No. 2, Feb. 1999, pp. 58–66.

[10] I. Cidon, Y. Ofek, "MetaRing – a Full-Duplex Ring with Fairness and Spatial Reuse," *IEEE Transactions on Communications*, Vol. 41, No. 1, Jan. 1993, pp. 110–120.

[11] A. Bianco, J.M. Finochietto, G. Galalnte, F. Neri, V. Sarra, "A Fairness Enforcement Protocol for Interconnected WDM Rings," ONDM 2004, 8th IFIP Working Conference on Optical Network Design and Modelling, Gent, Belgium, February 2-4, 2004.

[12] A. Bianco, M. Franceschinis, S. Ghisolfi, E. Leonardi, F. Neri, A.M. Hill, R. Webb, "Frame-Based Matching Algorithms for Input-Queued Switches," *IEEE Workshop on High Performance Switching and Routing (HPSR02),* Kobe, Japan, May 2002.