

Multi-Fasnet: a simple scheduling algorithm for a distributed optical switch

Andrea Bianco*, Elisabetta Carta[†], Davide Cuda*, Jorge M. Finochietto*, Fabio Neri*

* Dipartimento di Elettronica, Politecnico di Torino, 10129 Torino, Italy,
Email: {andrea.bianco, davide.cuda, jorge.finochietto, fabio.neri}@polito.it

[†] Master student at the Politecnico di Torino
Email: carta@studenti.polito.it

Abstract—The design of classic switching architectures for today’s telecommunication network needs to consider the limits imposed by the electronic technology; like, power supply, consumption and dissipation. The introduction of optical technology for switching functions can overcome the most of the current design limits. We propose a cost-effective architecture that implements optical switching without any need of optoelectronic conversion. Moreover, we proposed a distributed access scheme based on an extension of the Fasnet protocol, and we compare its performance with the classical switching algorithms.

I. INTRODUCTION

The last years have seen a continuous rise of bandwidth demand from many emerging networking applications, like VoIP, Video-On-Demand, Video-Conferencing and p2p. This rapid increase of traffic level is well supported by the optical technology, and in particular by the Wave-Length-Multiplexing (WDM) technique, which was able to assure significant improvements in network bandwidth. Optical technology has already emerged as the core transmission technology, due to its ability to carry huge amount of data traffic, but it is quite exclusively used to support point-to-point connections between nodes. Indeed, each node must perform optical-to-electrical conversion, and must electronically process the entire traffic for switching/routing. Consequently, one of the main concerns of the networking community, it is becoming the mismatch between the transmission capacity offered by the WDM optical layer and the processing capacity of the current routers/switches.

Today’s most common switch architectures are based on electronic fabrics to process data from input to output ports. Although electronic switching fabrics have scaled remarkably with the capacity demanded by routers, reaching aggregate capacity of few Tb/s; they require today to dissipate too much power and to occupy too much space to be practical. The simplest way to connect multiple ports is to use an electrically shared bus to handle interconnects, even if, nowadays, the preferred solution to realize fast switching fabrics is to use electrical crossbar fabric, since multiple electrical point-to-point connections can be setup between ports. However, to support an increasing number of ports or higher data rate, the clock frequency must increase leading to a larger power consumption and dissipation. Indeed, the more frequency increases, the more electromagnetic compatibility and power

density problems, as well the layout complexity, become the key limiting factor for the overall switch capacity; making these solutions unattractive for high speed switches.

Recently, the use of the optical technology for switching and more complex functions is rapidly gaining interest, both in the research and industrial communities. Besides the huge available bandwidth, the employment of the optical technology for switching presents very interesting aspects, like its reduced power consumption and dissipation, the larger possible distance, the great flexibility in designing and interconnecting different topologies but, most of all, the fact that the switching cost of wavelength is quite independent from the data bit-rate, in contrast to the electrical technology. On the other hand, the lack of optical memories makes very difficult to solve conflict in time domain through dynamic operations, which, actually, is the basis of the packet switching. For this reason a true optical packet switch seems very difficult to be implemented today.

Broadcast-and-Selected switching architectures, where packets are sent from any input ports to all outputs, and where each output port selects the data addressed to it, are an intermediate solution in-between fast optical circuit-switching and optical packet switching. In such architectures, packet switching is done only at the system edge, i.e., at the interface between the electrical and the optical domain, while packets are transmitted in a single-hop fashion on the optical domain where no contentions using network resources arise.

The paper is organized as follows. In Sec. II we describe the architecture under study, while in Sec. III we describe the well known Fasnet protocol and its adaptation to the proposed architecture. In Sec. IV we present simulation results of the proposed distributed access scheme, comparing its performance and its limitations to the classical centralized switching algorithms. Finally, we draw some conclusions and the guidelines for the future work in Sec. V.

II. SYSTEM MODEL

We consider a specific WDM optical packet switch, whose architecture was proposed, studied and prototyped in the framework of the Italian national project called OSATE [1]. The architecture of the OSATE optical switching fabric is depicted in Fig. 1, while the structure of a node is illustrated in Fig. 2. The OSATE architecture comprises N input ports

and N output ports connected by two counter-rotating WDM fiber rings. Each ring conveys W wavelengths, with $N \leq W$; in order to have a non-blocking switching matrix. Each ring is used in a specific way: one ring is used for transmission only, while the second ring is used for reception only. Transmission wavelengths are switched to the reception ring, at a folding point between the two rings, as shown in Fig. 1. During the first ring traversal, transmitted packets cross the transmission ring until the folding point, where they are switched to the reception ring and then received during the second ring traversal. As such, the architecture behaves as a folded bus network.

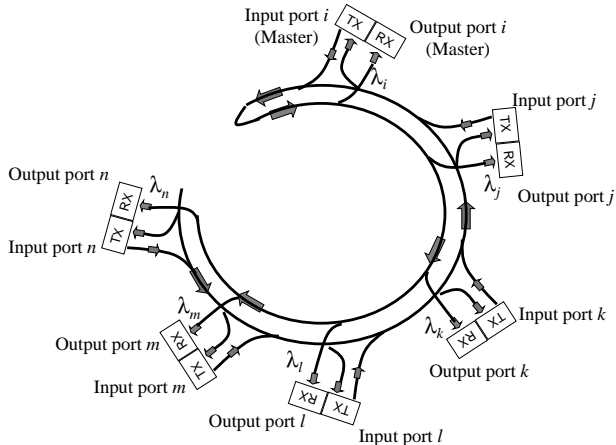


Fig. 1. OSATE switching fabric architecture

The network is synchronous and time-slotted. The slot duration is determined by technological constraints, such as tuning time and dispersion, by user packet sizes, and by the efficiency of the packet segmentation process. We take $1\mu s$ as a reference value for the slot duration. During a time slot, at most one packet can be transmitted by an input port in one of the W available slots (one slot for each wavelength channel). Each input is equipped with a fixed receiver, tuned to λ_{drop} in Fig. 2; given that $N \leq W$, each output port is allocated to a single WDM channel, as described in [4]. To provide full connectivity between nodes, each input port is equipped with a *fastly* tunable transmitter (implemented as an array of fixed lasers, as shown in Fig. 2) and exploits WDM to partition the traffic directed to each destination output. Input ports tune their transmitters to the receiver's destination wavelength, establishing a single hop connection lasting one time slot. The channel resource sharing is therefore achieved according to a Time Division Multiple Access (TDMA) scheme. Moreover, to avoid Head of the Line (HoL) [6] problem, each node is equipped with W queues; indeed, if a single FIFO is used, a packet at the head the queue might block other packets which could be transmitted on others channels.

A collision may arise when an input tries to insert a packet on a time slot and wavelength which have already been used. Thus, access decisions are based on channel inspection capability (similar to the carrier sense functionality in Ethernet), called λ -monitor. In this way, each input port knows which wavelengths have not been used by upstream inputs in the

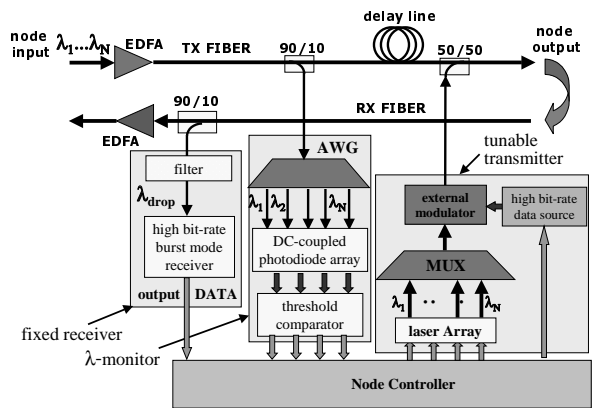


Fig. 2. OSATE node structure

current time slot. Priority is given to in-transit traffic, i.e., a *multi-channel empty-slot* protocol is used.

Packets traversing the proposed optical backplane might suffer collisions if a proper scheduling algorithm is not defined. Moreover, the ring topology of the switching fabric impose that input ports transmit packets sequentially and not in parallel like in the traditional crossbars, indeed, a simple empty slot scheme might lead to fairness problem due to different access probability depending on the position of the input ports along the ring. Referring to Fig. 1, an upstream input can “flood” a given wavelength, as shown in [8], reducing (or even blocking) the transmission opportunities of downstream ports competing for access to that channel, thus leading to significant fairness problems. Therefore, a suitable distributed switching algorithm must be able to assure large level of throughput, bounded delay and equal transmission probability even when the inputs are heavy loaded.

Although distributed algorithm are usually simpler to implement, requiring little or no control information exchange, they might show limited performance. On the contrary, centralized schemes are usual more complex and require heavy information exchange, but can easily schedule packets and achieve large level of throughput. In the following we first describe the distributed Multi-Fasnet access scheme, then its performance are compared with respect to centralized switching algorithms, like iSLIP [7] and Maximum Weight Matching [6].

III. THE FASNET PROTOCOL

Fasnet [10] is an access protocol originally designed to guarantee fairness on a slotted dual bus topology. In the following subsections, we first analyze the protocol in a folded bus topology with a single channel; next, we adapt the protocol for a multichannel architecture like OSATE.

Fasnet is an implicit token passing protocol developed to efficiently use the channel capacity, providing a high level of fairness in resource sharing. To implement Fasnet, all nodes should listen on the transmission channel, excluding the first node in the bus, dubbed master node, which has to listen on the reception one. As shown in Fig. 2, all nodes are equipped

with a λ -monitor that allows them sensing the transmission channel, but not the reception one. However, this can be easily implemented by simply giving to each node the possibility to switch its own λ -monitor between the transmission bus and the reception one. In fact, the master node, being the first node on the transmission bus, does not experiment any packet collisions on the transmission bus, so it can switch the λ -monitor to the reception channel, while, all the other nodes can switch it to the transmission one.

Fasnet provides fairness operating cyclically; each cycle is associated with a chained transmission of data called train. A train is composed by a first packet, dubbed locomotive, transmitted by the master node, and by all packets transmitted by network nodes after the locomotive. The master node starts a new cycle, transmitting a new locomotive, every time it detects the end of the in-transit train (i.e., an empty slot on the reception channel). Each node is assigned a quota Q , which represents the maximum number of packets that can be transmitted when an empty slot after a locomotive is detected. When a node senses an end of train, it seizes the channel for a number of packets equal to the minimum between the quota Q and the number of packets in its queue. Once a node releases the channel (either by exhausted quota or empty queue), it restores its quota and waits for the next train before attempting to access the channel again.

Note that Fasnet is not able to reach 100% throughput, due to the idle time between two successive cycles. Indeed, the master node recognize the end of train only when the last transmitted packet is sensed on its λ -monitor on the reception channel; this implies that a new locomotive is sent when no packets are traveling in the network. Thus, the maximum achievable throughput, when the network is overloaded, is mainly affected by the ratio between the maximum train length, which is equal to $N \times Q$ and the cycle duration, which is equal to $N \times Q$ plus the time needed by the master node to detect the end of the current train. In the OSATE architecture, this idle time is approximately twice the ring propagation delay, named round trip time (RTT) in the paper; during this time all transmitters remain idle. This implies that the maximum achievable throughput under uniform traffic is given by:

$$TH_{max} = \frac{N \times Q}{N \times Q + [2 \times RTT] + 1} \quad (1)$$

As a result, the larger the value of Q , the larger the maximum achievable throughput.

If we assume that the network is not overloaded, which means that a node empties its queue without exhausting its quota, we can easily estimate the worst case access delay. This happens when a packet arrives as soon as the node has just released the channel; the node has to wait for the next train to transmit this packet. Therefore, the worst case access delay at low loads can be evaluated as:

$$D_{WC} \approx N \times Q^* + [2 \times RTT] + 1 \quad (2)$$

where Q^* is the effective average quota used by a node. Q^* can be evaluated considering that, under lightly loaded conditions, the throughput TH is equal to the input load ρ . Therefore, from (1) we obtain:

$$Q^* = \frac{\rho}{1 - \rho} \times \frac{[2 \times RTT] + 1}{N} \quad (3)$$

Observe that Q^* , at low loads, does not depend on the value of Q , but is a function of the input load and the network dimension; indeed, the train length adapts to the network load.

The performance of the Fasnet protocol is limited both in throughput and in delay by the channel idle time needed by the master node to detect the end of the current cycle. We are in front of a trade-off: on the one hand, we want a large value of quota to achieve high throughput but, on the other hand, if we want to ensure low access delays, a low quota value is needed.

A. Multi-Fasnet Protocol

In a multichannel network, the Fasnet behavior is replicated over the different wavelengths, which means that there are W trains, one for each channel, traveling across the network. If, in the same time slot, a node can access more than one channel, then a *train collision* happens. Since nodes are equipped with a single fastly tunable transmitter (see Fig. 2), they can transmit at most one packet per time slot. Thus, when a train collision occurs, nodes select the channel to which the longest queue is associated to. This means that nodes may release a channel although they still have both quota and packets to transmit simply because a train collision occurred. If this is the case, nodes are allowed to transmit on the next cycle at most Q packets plus the remaining quota of the previous cycle. In this way, if train collisions happen, fairness can be still reached in more than one cycle. To avoid excessive quota accumulation, the maximum quota that can be accumulated on a channel is bounded by either the node current queue length on the corresponding channel, or by $M \times Q$, where M is a parameter set to 5 in simulation experiments.

To estimate the maximum throughput in a multichannel network for Bernoulli traffic, we need to take into account the traffic matrix; (1) becomes:

$$TH_{max} = \frac{1}{W} \times \sum_{w=1}^W \frac{\sum_{i=1}^N \lambda_{iw} \times Q}{\sum_{i=1}^N \lambda_{iw} \times Q + [2 \times RTT] + 1} \quad (4)$$

where λ_{iw} is the average traffic sent by node i on channel w .

The worst case access delay on wavelength w at low loads becomes:

$$D_{WC_w} \approx \sum_{i=1}^N \lambda_{iw} \times Q_{iw}^* + [2 \times RTT] + 1 \quad (5)$$

where Q_{iw}^* is the effective average quota used by node i on channel w .

Therefore, Multi-Fasnet performance is also limited by the channel idle time in a multichannel network. To improve Multi-Fasnet performance, we must reduce the fixed penalty of having an idle channel for $2 \times RTT$ slots between two cycles. Thus, the master node must start a new train without waiting to sense the end of the current train.

IV. PERFORMANCE EVALUATION

We present performance results obtained by simulation considering a network with $W = 16$ wavelengths and a total of $N = 16$ nodes. The inter-node distance is about 100 ns and each node introduces a delay of 100 ns to perform the void detection; thus, the RTT of each ring is equal to $3.1 \mu\text{s}$. Each node keeps W separate FIFO queues, one for each channel, with a queue size of about 32000, fixed size, packets.

We compare the Multi-Fasnet access strategy with classic scheduling algorithms, like iSLIP and Maximum Weight Matching (MWM). Different traffic scenario are considered: uniform traffic and unbalanced traffic. To describe the traffic scenarios, let be ρ_i is the load at input port i , and λ_{ij} the load from the input port i to the output port j .

In the uniform traffic, the whole network capacity is equally shared by all nodes, i.e., each input port transmits with probability $\lambda_{ij} = \rho_i \times 1/N$ to each output. Two different unbalanced traffic patterns are considered: the bi-diagonal traffic and the log-diagonal traffic. In the bi-diagonal traffic each input port i transmits to an output port j according the following rates:

$$\lambda_{ij} = \begin{cases} \rho_i \times \frac{2}{|i-j|+1} & \text{if } j = i \\ \rho_i \times \frac{1}{|i-j|+1} & \text{if } j = |i+1|_N \\ 0 & \text{otherwise} \end{cases} \quad (6)$$

where $|x|_N = x \bmod N$. This input load is skewed since node i only has traffic for output port i and $i+1 \bmod N$. For log-diagonal traffic scenario, $\lambda_{ij} = 2 \times \lambda_{|i+1|_N}$ and $\sum_j \lambda_{ij} = \rho_i$.

We mainly focus on delay-throughput plots, obtained by simulation.

Fig. 3 shows the performance of the Multi-Fasnet, iSLIP and MWM algorithms under uniform traffic pattern. MWM and iSLIP achieve 100% throughput, while the Multi-Fasnet performance are dramatically affected by the value of the quota. As discussed Sec. III-A, the larger is the quota the larger is the network utilization, since the idle time between the two consecutive cycles has a lower impact as the train length increases. The maximum achievable throughput evaluated using (4) is, respectively, $TH_{max} = 0.67$ for $Q = 1$, $TH_{max} = 0.95$ for $Q = 100$ and $TH_{max} = 0.995$ for $Q = 100$, a little larger than the values obtained by simulation; this little difference is mainly due to the train collision effect. When the network is lightly loaded, the mean transmission delay is independent of the quota value; indeed, the train length depends on input traffic and network dimension only. With respect to a centralized scheme, the Multi-Fasnet protocol shows a quite constant larger transmission delay equal to

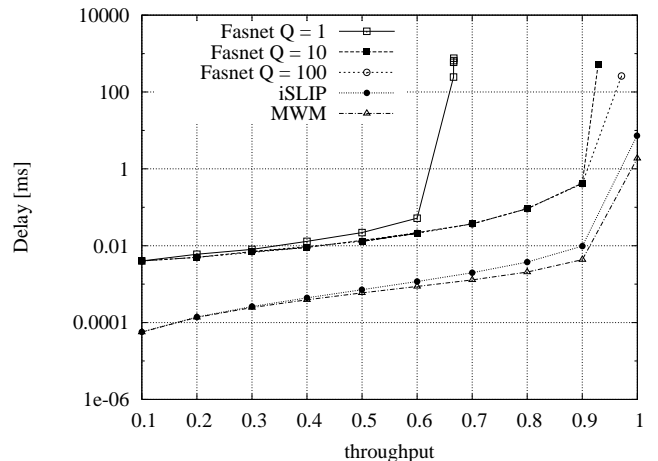


Fig. 3. Multi-Fasnet, iSLIP and MWM performance under uniform traffic scenario

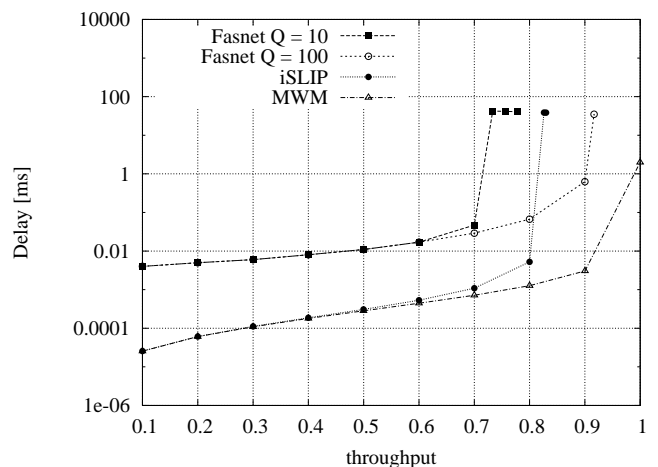


Fig. 4. Multi-Fasnet, iSLIP and MWM performance under log-diagonal traffic scenario

$2 \times RTT \mu\text{s}$; indeed, the network remains idle for $2 \times RTT$ slots between two cycle (like explained in Sec. III, $2 \times RTT$ is the time the master node needs to detect the end of the in-transit train and to start a new cycle). In overloaded conditions, the differences between a centralized scheme and a distributed one drastically decrease, since the mean delay depends on the access delay plus the time needed to traverse the whole queue length QL . Under uniform traffic, in overload, all nodes access the channel after $D_{WC_k} = N \times Q + [2 \times RTT] + 1 \mu\text{s}$ (slots) and transmit Q packets: the mean delay is equal to $D_{WC_k}/Q \times QL \mu\text{s}$. Thus, the mean delay in overloaded conditions is approximately equal to 768 ms for $Q = 1$, 537 ms $Q = 10$ and 514 ms $Q = 100$.

We conclude this Sec. comparing the Multi-Fasnet with respect to iSLIP and MWM performance under two unbalance traffic scenarios. Fig. 4 and Fig. 5 show the throughput versus delay plots under the bi-diagonal and the log-diagonal traffic scenarios, respectively. Although the Multi-Fasnet protocol always shows a larger delay when the network is lightly loaded due to the idle time between two trains, as the network load increases, the differences between iSLIP and the Multi-Fasnet

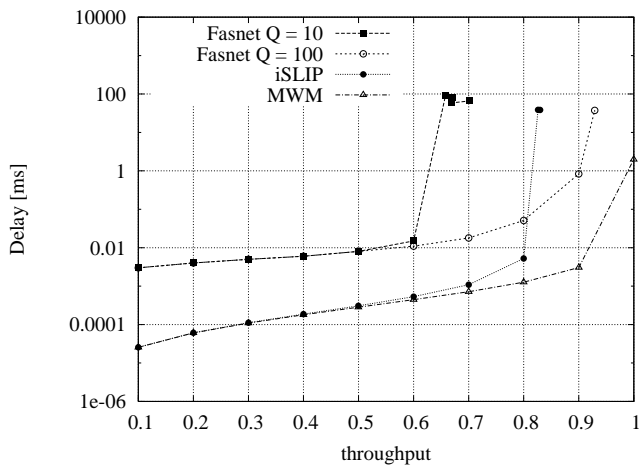


Fig. 5. Multi-Fasnet, iSLIP and MWM performance under bi-diagonal traffic scenario

protocol decreases, and with a large enough quota, the Multi-Fasnet protocol is able to achieve a larger throughput than iSLIP.

V. CONCLUSIONS AND FUTURE WORK

We introduced a particular WDM ring-based distributed switching fabric called OSATE, and discussed the Multi-Fasnet strategy, the adaptation of an existing protocol to this WDM scenario, and we compared it with the well-known performance of centralized switch.

Simulation results show how Multi-Fasnet performance are mainly limited by the channel idle time between two consecutive cycles; thus, Multi-Fasnet needs large value of quota to reach large throughput. Although these limitations, the Multi-Fasnet protocol achieves large network utilization and it is very simple to be implemented in optical distributed switch, where the computing power is very limited.

We aim to continue our work improving the Multi-Fasnet protocol, reducing the impact of the network dimensions on the node access delay. The main idea is to retransmit a train without that the master node must wait until it detects the end of the in-transit train. Moreover, since it is well known fact, that both the complexity of centralized scheme remarkably increases and performance decrease (see [9]), when centralized algorithms are implemented over a distributed architecture; we aim to study how classic switching algorithms, like iSLIP, must be modified and which are their performance when they are implemented over the proposed network architecture.

ACKNOWLEDGMENT

This work was performed in the framework of the FP6 European Network of Excellence e-Photon/ONE.

REFERENCES

[1] The OSATE Project: <http://www.tlc-networks.polito.it/projects/osate/>
 [2] The WONDER Project: <http://www.tlc-networks.polito.it/wonder/>
 [3] A. Carena, V. De Feo, J. M. Finochietto, R. Gaudino, F. Neri, C. Piglione, P. Poggiolini, "RingO: An Experimental WDM Optical Packet Network for Metro Applications," *IEEE Journal on Selected Areas in Communications*, vol. 22, no. 8, , pp. 1561-1571, Oct. 2004

[4] A. Bianco, J. M. Finochietto, G. Giarratana, F. Neri, C. Piglione, "Measurement Based Reconfiguration in Optical Ring Metro Networks", *IEEE/OSA Journal of Lightwave Technology (JLT)*, Special Issue on "Optical Networks", vol.23, no.10, pp.3156-3166, October 2005
 [5] A. Bianco, E. Di Stefano, A. Fumagalli, E. Leonardi, F. Neri, "A Posteriori Access Strategies in All-Optical Slotted Rings", *IEEE GLOBECOM'98*, November 1998, Sydney, Australia
 [6] N. McKeown, A. Mekkittikul, V. Anantharam, and J. Walrand, "Achieving 100% throughput in an input-queued switch", *IEEE Transactions on Communications*, Vol. 47, No. 8, pp. 1260-1267, Aug. 1999.
 [7] N. McKeown, "The iSLIP Scheduling Algorithm for Input-Queued Switches", *IEEE/ACM TRANSACTIONS ON NETWORKING*, Vol. 7, No. 2, April 1999.
 [8] M. Ajmone Marsan, A. Bianco, E. Leonardi, M. Meo, and F. Neri, "MAC Protocols and Fairness Control in WDM Multi-Rings with Tunable Transmitters and Fixed Receivers," *IEEE/OSA Journal on Lightwave Technology*, Vol. 14, No. 6, pp. 1230-1244, Jun. 1996.
 [9] C. Minkenberg, F. Abel, E. Schiattarella, "Distributed crossbar schedulers", *HPSR*, High Performance Switching and Routing, June 2006.
 [10] J. O. Limb and C. Flores "Description of Fasnet – A Unidirectional Local-Area Communication Network", *The Bell System Technical Journal*, Vol. 61, No. 7, September 1982.
 [11] A. Bianco, J.M.Finochietto, G.Galante, F.Neri, V.Sarra, "Scheduling Variable-Size Packets in the DAVID Metropolitan Area Network", *IEEE ICC 2004 (IEEE International Conference on Communications)*, June 2004, Paris, France