Multi-Fasnet Protocol: Short-Term Fairness Control in WDM Slotted MANs

Multi-Fasnet protocol: short-term fairness control in WDM slotted MANs

Andrea Bianco*, Davide Cuda*, Jorge M. Finochietto*, Fabio Neri*, and Chiara Piglione†
* Dipartimento di Elettronica, Politecnico di Torino, 10129 Torino, Italy,
Email: {andrea.bianco, davide.cuda, jorge.finochietto, fabio.neri}@polito.it
† DCBU Modelling and Simulation Team, CISCO Systems, San José, USA ,
Email: cpiglione@cisco.com

***Abstract*— Single-hop wavelength division multiplexing (WDM) ring networks operating in packet mode are a promising architecture for the design of innovate Metropolitan Area Networks. They allow a cost-effective design, with a good combination of optical and electronic technologies, while supporting features like *restoration* and *reconfiguration* that are essential in any metropolitan scenario. In this article, we address the fairness problem in a slotted WDM optical network. We introduce the Multi-Fasnet fairness protocol, we discuss its limitations and we propose an extension, based on a dynamic strategy, that achieves high aggregate network throughput, throughput fairness, and bounded and fair access delays.**

I. Introduction

Optical packet switching (OPS) architectures are excellent candidates to meet the requirements of future MANs. Since truly header-based packet switching in the optical domain is not mature today, we focus on a WDM broadcast-and-select solution, in which transmitted packets reach their destination in a single all-optical hop. We further concentrate on a ring topology, due to its well known fault recovery properties.

The choice of a broadcast and select optical packet ring network leads to the problem of designing an efficient protocol to arbitrate the access of nodes to channel resources. An efficient access protocol must be able to optimize network throughput, controlling under which conditions packets stored electronically in nodes can be transmitted. The challenge is to obtain high network utilization while minimizing delay and providing acceptable efficiency/fairness trade-offs.

We extend the Fasnet protocol, originally proposed for electronic networks, to a multi-channel environment; we highlight its limitations and propose a novel strategy able to dynamically adapt the fairness mechanism to the traffic pattern without requiring complex measurements procedure. The strategy is simple to implement, yet provides high throughput, bounded delays and shows good fairness properties.

II. System Model

We consider a specific WDM optical packet network, whose architecture was proposed, studied and prototyped in the framework of the Italian national project called WONDER [1]. The architecture of the WONDER network [2] is depicted in Fig. 1, while the structure of a node is illustrated in Fig. 2. The WONDER architecture comprises N nodes connected to two counter-rotating WDM fiber rings. Each ring conveys W wavelengths, with $N > W$; 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.

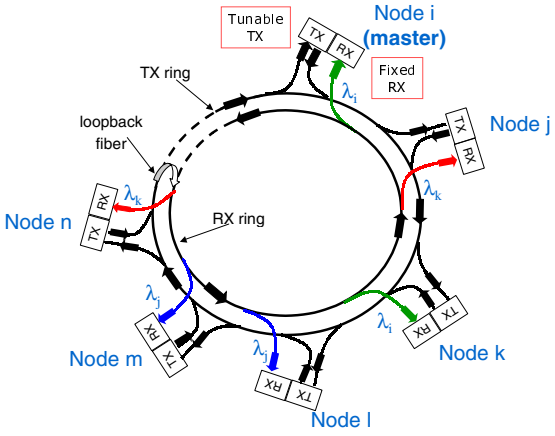

Fig. 1. WONDER network architecture

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 1μs as a reference value for the slot duration. During a time slot, at most one packet can be transmitted by a node in one of the W available slots (one slot for each wavelength channel). Each node is equipped with a fixed receiver, tuned to λ_{drop} in Fig. 2; given that $N > W$, more than one node receives from a given wavelength channel. Receivers are allocated to WDM channels in a way that equalize the traffic across WDM channels, as described in [3]. To provide full connectivity between nodes, each node 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

1-4244-0357-X/06/$20.00 ©2006 IEEE
This full text paper was peer reviewed at the direction of IEEE Communications Society subject matter experts for publication in the IEEE GLOBECOM 2006 proceedings.

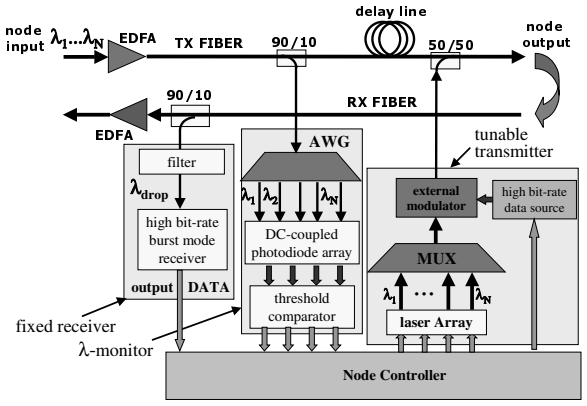

Fig. 2. WONDER node structure

to disjoint subsets of destination nodes; each subset is made by the nodes whose receivers are currently allocated on the same wavelength. Nodes 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.

A collision may arise when a node 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 node knows which wavelengths have not been used by upstream nodes in the current time slot. Priority is given to in-transit traffic, i.e., a *multi-channel empty-slot* protocol is used.

III. The Fairness Problem

From a design perspective, a suitable access protocol for the WONDER network must be adaptable to a multi-channel network, not only avoiding packet collisions but also assuring some level of fairness together with acceptable network throughput and delays bounds. A first level of fairness is achieved implementing an efficient a-posteriori [4] packet selection strategy and exploiting the Virtual Output Queue (VOQ) structure [5]. The basic VOQ idea consists in using separate queues, each one corresponding to a different destination, or to a different set of destinations (e.g. all the nodes receiving on the same wavelength), and to appropriately select the queue which gains access to the channel for each time slot. Indeed, the WONDER network can be seen as a distributed IQ switch, where the ring plays the role of the switching fabric. Since, in general, there is more than one node receiving on the same wavelength, there is no difference between adopting a queue for each destination (N queues) or a queue for each channel (W queues); for simplicity and without loss of generality we adopt the second solution.

A problem common to ring and bus topologies is the different access priority given to network nodes depending on their position along the ring/bus. Referring to Fig. 1, an upstream node can "flood" a given wavelength, as shown in [6], reducing (or even blocking) the transmission opportunities of downstream nodes competing for access to that channel, thus leading to significant fairness problems.

Another issue is that, due to the application to MANs, the end-to-end propagation delay is fairly larger than the average packet transmission time. For example, a 50 km span leads to a 250 μs propagation delay, while a 1000-bit packet at 1Gbit/s lasts 1 μs. This makes it difficult for a distributed fairness control scheme to act on short time scale, (comparable with packet duration) and to keep access times within reasonable low bounds to support time critical applications.

IV. The Fasnet Protocol

Fasnet [7] 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 network like WONDER, and conclude proposing some new, more efficient, strategies.

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

1-4244-0357-X/06/$20.00 ©2006 IEEE
This full text paper was peer reviewed at the direction of IEEE Communications Society subject matter experts for publication in the IEEE GLOBECOM 2006 proceedings.

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 WONDER 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:

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

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 the evaluated as:

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

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:

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

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:

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

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:

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

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.

B. Fixed-Length Train Strategy

The first strategy is called the Fixed-Length Train (FLT). On a given channel k, $k = 1, \ldots, W$, the master node has the possibility to schedule a new train every C_k time slots, where C_k is a counter initialized to $N \times Q$ each time a new train is transmitted on channel k. Since C_k is decreased by one at each time slot, using the FLT strategy a new train is scheduled if one of the following events happen:

- if the master node senses an end of train (as in the original Fasnet protocol), and there are no other trains propagating along the channel. This condition may happen only if C_k is initialized to a value larger than $2 \times RTT$;
- if $C_k = 0$, since the largest possible train length has been reached.

When using the FLT strategy, nodes access channels cyclically, like in a Time Division Multiplexing (TDM) scheme. The advantage of the FLT technique consist in the fact that, when the network is overloaded, no slots are left empty unless when train collision occurs. Problems might arise if the train length does not match the traffic scenario (e.g. under non uniform traffic patterns); in this case the trains are left partially empty and throughput losses are experienced.

1-4244-0357-X/06/$20.00 ©2006 IEEE
This full text paper was peer reviewed at the direction of IEEE Communications Society subject matter experts for publication in the IEEE GLOBECOM 2006 proceedings.

C. Dynamic-Length Train Strategy

A fixed length train strategy is not efficient for variable traffic patterns. The idea of the Dynamic-Length Train (DLT) strategy is to estimate train lengths taking into account the current traffic load, without waiting to sense an end of train at the master node. The DLT strategy tries to determine the "optimal" train length by looking at the train utilization, i.e., the percentage of used slots of the last received train. Indeed, if trains are partially empty, then the train length can be decreased, augmenting its scheduling frequency. On the contrary, if trains are completely full, their length should be increased, lowering the train scheduling frequency.

Also the DLT mechanism allows the master node to start a new train every C_k slots; but, unlike the FLT strategy, the value of C_k is not constant but variable and estimated by traffic measurements. The value of C_k is updated every time the master node detects the end of a train by considering the train utilization, i.e., the number of busy slots on the train. In particular, C_k is updated in the following way:

$$C_k = \begin{cases} C'_k + I \times C'_k & \text{iff all train slots are busy} \\ C'_k - D \times C'_k & \text{otherwise} \end{cases}$$

where C'_k refers to the value of C_k in the previous cycle, and I and D are two parameters that denote the increase and decrease steps, set respectively to 0.3 and 0.1 in our simulations.

V. Performance Evaluation

We present performance results obtained by simulation considering a network with $W = 4$ wavelengths and a total of $N = 16$ nodes, for a total ring length of about 25km. Slots last 1μs, corresponding to a fixed packet size of about 1250 bytes at 10 Gbit/s; thus, the ring RTT is equal to 121 slots. Each node keeps W separate FIFO queues, one for each channel, with a queue size of about 120000, fixed size, packets.

Two different traffic scenarios are considered: uniform traffic and unbalanced traffic. In the uniform traffic pattern, the whole capacity of the network is equally shared by all nodes. In the unbalanced traffic pattern, named "1-server", nodes are partitioned into two separated subsets: server $\mathcal{S}$ and clients $\mathcal{C}$. The server subset contains only a single node, named *server*, positioned at the head of the bus to provide a worst case scenario. The server transmits at a high rate, equal to the capacity of one wavelength, with equal probability to the other $N-1$ nodes belonging to $\mathcal{C}$. The remaining network capacity is shared by client nodes; each client transmits $\frac{1}{3}$ of its traffic toward the server and the remaining traffic to the other $N-2$ clients with equal probability.

We mainly focus on delay-throughput plots obtained by simulation; fairness is evaluated as the difference between the performance achieved by the first and the last node on the bus.

We first consider, in Fig. 3, protocol behavior under uniform traffic pattern with two different values of quota: $Q = 10$ and $Q = 100$. The Multi-Fasnet maximum throughput is dramatically affected by the value of the quota. As discussed in Section IV-A, the larger the quota, the larger the network utilization, since the idle time between two consecutive cycles has a lower impact if the train length increases. The maximum achievable throughput evaluated using (4) is, respectively, $TH_{max} = 0.4$ for $Q = 10$ and $TH_{max} = 0.87$ for $Q = 100$, correctly matched by simulation. 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. However, in overloaded conditions 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 $D_{WC_k} = N \times Q + 2 \times RTT$ μs (slots) and transmit Q packets: the mean delay is equal to $D_{WC_k}/Q \times QL$ μs. Thus, the mean delay in overloaded conditions is approximately equal to 4800 ms for $Q = 10$ and 2210 ms $Q = 100$.

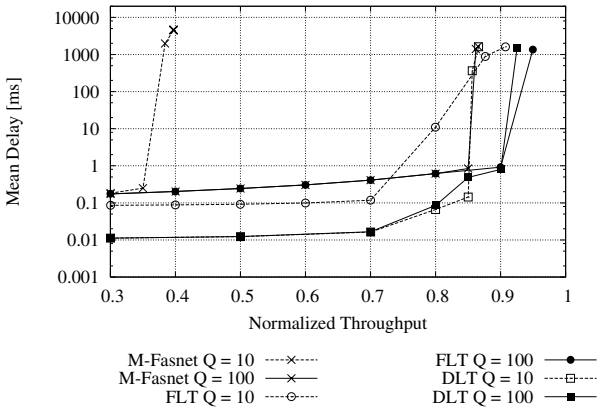

Fig. 3. Multi-Fasnet, FLT and DLT performance under uniform traffic scenario for $Q = 10$ and $Q = 100$

Let us now focus on the FLT and DLT strategies under uniform traffic scenario; both strategies improve the network utilization since they are able to cope with the RTT induced idle time. Under uniform traffic scenario, the FLT train length is matched to the traffic pattern; thus, a high throughput can be achieved. Some throughput losses are induced by the train synchronization effect. The DLT strategy presents some throughput losses with respect to the FLT strategy; the trains are left partially empty since their average length is larger than the optimal one (equal to $N \times Q$ under uniform traffic pattern). Hoverer, the DLT strategy is able to reduce the transmission delay of about one order of magnitude when the network is lightly loaded. Indeed, when the network is lightly loaded, the trains are left partially empty. According to the DLT strategy rules, the master node increases the train generation frequency, reducing the train length; all nodes can access the channels more frequently, thus decreasing their mean delay.

Note that both FLT and DLT are significantly more robust to the quota value chosen, a positive effect provided by these strategies, although increasing Q still provides some benefit. FLT suffers for higher delays at medium loads. Throughput fairness (not shown) is very good for all strategies, including Multi-Fasnet.

Let us compare Multi-Fasnet, the FLT strategy and the DLT strategy under a 1-server traffic scenario. The network

1-4244-0357-X/06/$20.00 ©2006 IEEE
This full text paper was peer reviewed at the direction of IEEE Communications Society subject matter experts for publication in the IEEE GLOBECOM 2006 proceedings.

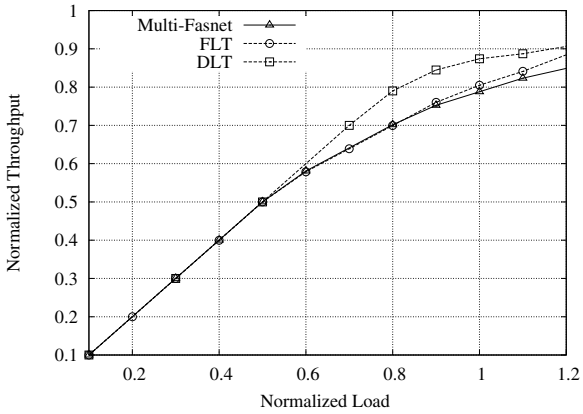

Fig. 4. Comparison between Multi-Fasnet, the FLT strategy and the DLT strategy throughput under 1-server scenario ($Q = 100$)

throughput is plotted against the offered load in Fig. 4. Although the FLT strategy is very efficient under uniform traffic, its performance are quite limited under unbalanced traffic conditions, as expected. The Multi-Fasnet protocol suffers from idle times between trains, that limit the maximum throughput. On the contrary, the DLT strategy is able to match the train length to the traffic pattern, thus achieving a larger throughput.

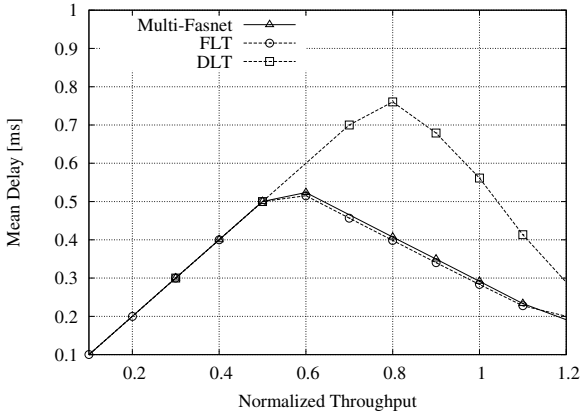

Fig. 5. Comparison of the server performance in Multi-Fasnet, the FLT strategy and the DLT ($Q = 100$).

Performance increase is even more visible if we analyze server throughput, plotted in Fig. 5. Adopting the DLT strategy, the server is able to achieve larger value of throughput (about 0.75) before being starved. Clearly, when the network is overloaded, the throughput of all the nodes converges to the same value according to the max-min fairness paradigm.

Delay fairness performance is shown in Fig. 6 for Multi-Fasnet and DLT strategy under uniform traffic, which is the more critical scenario for delays. As expected, Multi-Fasnet provides a very high level of fairness in terms of delays; indeed, all nodes have the same access probability in each cycle. The DLT strategy maintains the very good fairness level achieved by the Multi-Fasnet protocol, and reduces the access delay significantly at low loads. However, it presents some delay unfairness when the network in lightly loaded. This is mainly due to the fact that the last nodes might lose some

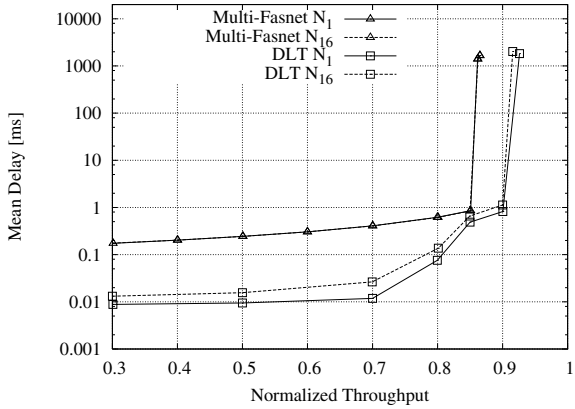

Fig. 6. Multi-Fasnet and DLT strategy fairness under uniform traffic scenario ($Q = 100$)

cycles when the train length is underestimated.

VI. Conclusions and Future Work

We introduced the fairness problem in a particular WDM ring-based OPS network architecture named WONDER, and discussed Multi-Fasnet, the adaptation of an existing protocol to this WDM scenario. Moreover, we proposed two new fairness strategies, named FLT and DLT.

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 high throughput, leading to large access delays. The simple FLT strategy is limited by the need to match the train length to the traffic scenario. The DLT strategy shows robustness to parameter setting, high throughput, low acess delays and good fairness properties and seems a promising solution to the fairness issue in WDM ring-based OPS networks.

Acknowledgment

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

References

[1] The WONDER Project: http://www.tlc-networks.polito.it/wonder/

[2] 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

[3] 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

[4] 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

[5] 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.

[6] 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.

[7] 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.