A Distributed Scheduling Algorithm for an Optical Switching Fabric

This full text paper was peer reviewed at the direction of IEEE Communications Society subject matter experts for publication in the ICC 2008 proceedings.

A distributed scheduling algorithm for an optical switching fabric

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, elisabetta.carta@gmail.com
† CONICET - Universidad Nacional de Cordoba, Argentina, Email: {jorge.finochietto}@ieee.org

***Abstract*— Designing switching architectures for network routers and switches needs to consider limits imposed by the electronic technology, like small bandwidth×distance factors, power density constraints, energy consumption and dissipation issues. Introducing optical technologies to implement switching functions can overcome several of the current design limits. We propose a cost-effective architecture implementing an optical switch without any need for optoelectronic conversion within the switching fabric. We further propose a distributed scheduling scheme, based on an extension of the Fasnet protocol, and we compare it to classical centralized scheduling algorithms, showing that a distributed scheduler can provide performance comparable to the ones offered by more complex centralized schedulers.**

I. Introduction

Many emerging networking applications, like VoIP, Video-On-Demand, Video-Conferencing and peer-to-peer, demand more and more communication bandwidth. Optical technologies, in particular by Wavelength Division multiplexing (WDM) techniques, have already emerged as the winning approach in transmission systems, due to the ability to transport a huge amount of information over large distances; still, their use is confined to support point-to-point connections between network nodes. Indeed, each switching node must perform optical-to-electrical conversion, electronically processing all the traffic for switching/routing. Consequently, the networking community is focusing its attention on the mismatch between the transmission capacity offered by the WDM optical layer and the processing capacity of current routers/switches.

Most common switching architectures are based on electronic crossbars to transfer data among different input/output ports. Electronic switching fabrics have scaled remarkably and can still keep up with the capacity currently demanded by routers, reaching today aggregate capacities up to a few Tb/s; nevertheless, they have almost reached their limits. Indeed, to support an increasing number of ports and higher data rates, the clock frequency must increase, leading to larger power consumption and dissipation issues. However, by increasing the operational frequency, electromagnetic compatibility and power density problems, as well as the layout complexity, become the key limiting factors for the overall switch capacity. Thus, these solutions are unattractive for future high-speed switches. A second major limit on the performance of packet switches is given by switching control algorithms for contention resolution and QoS enforcement: running these algorithms in a centralized manner introduces processing power and latency problems.

The use of optical technologies for switching is gaining interest, both in the research and industrial communities. Indeed, employing optical technologies for switching presents interesting aspects: huge available bandwidth, reduced power consumption and dissipation, intrinsic flexibility in supporting different interconnection topologies and especially, a wavelength switching cost quite independent from the data bit-rate (differently from the electronic domain). Despite all the advantages, implementing a fully optical packet switch is far from being convenient. Indeed, the lack of optical memories and the very limited processing capabilities in the optical domain, make it very difficult to solve conflicts in time domain through dynamic operations, which are indeed the basis of the packet switching concept.

Broadcast-and-select switching architectures, where packets are sent from any input port to all outputs, and where each output port selects the data addressed to it, represent an intermediate solution 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 arise. We present an optical architecture based on these principles, and suited to build an optical switching fabric. A prototype of this architecture is up and running in our labs and is now under testing.

To avoid the performance penalties due to centralized switch control schemes, a distributed scheduling algorithm, based on the Fasnet protocol, is proposed. In [1], we extended Fasnet in the context of WDM Metropolitan Area Networks, where large propagation delays are involved, so those solutions are different from the ones proposed in this paper. Finally, we also compare the submitted distributed switch control scheme with traditional centralized schedulers. Despite its simplicity, Fasnet shows performance close to those achieved by centralized schedulers in the considered setup.

The paper is organized as follows. In Sec. II we describe the architecture under study, in Sec. III we describe how the Fasnet protocol adapted to the proposed architecture. In Sec. IV we compare by simulation the proposed distributed access scheme with classical centralized scheduling algorithms. Finally, we draw some conclusions and guidelines for future work in Sec. V.

978-1-4244-2075-9/08/$25.00 ©2008 IEEE

II. System Model

Our WDM optical packet switch architecture was proposed, studied and prototyped in the framework of the Italian national project called OSATE [2]. The architecture of the OSATE optical switching fabric is depicted in Fig. 1, while the structure of a switch port 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. Rings are used in a peculiar 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, reach the folding point, are switched to the reception ring and finally received during the second ring traversal. As such, the architecture behaves as a folded bus topology, but it inherits the fault recovery properties of rings.

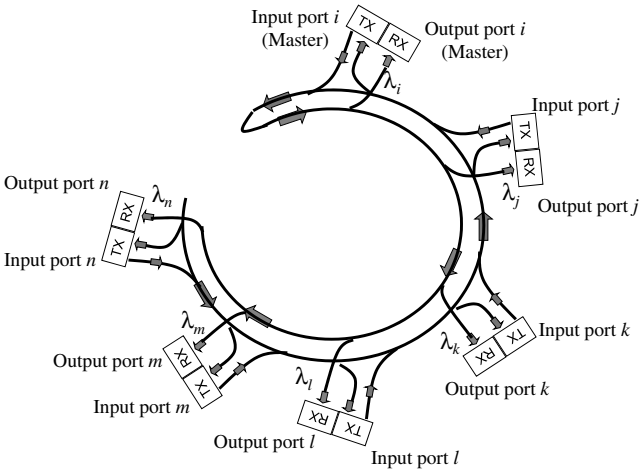

Fig. 1. OSATE switching fabric architecture

The switching fabric is synchronous and time-slotted. The slot duration is determined by technological constraints, such as tuning times and dispersion, by user packet sizes, and by the efficiency of the packet segmentation process. We take 1 μs as a reference value for the slot duration. Each port is equipped with a fixed receiver, tuned to λ_{drop} in Fig. 2; hence each output port is allocated to a single WDM channel. To provide full connectivity between ports, each input port is equipped with a *fastly* tunable transmitter (implemented as an array of fixed lasers in our prototype, as shown in Fig. 2 – see also [3]), and exploits WDM to partition the traffic directed to different destination ports.

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). 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.

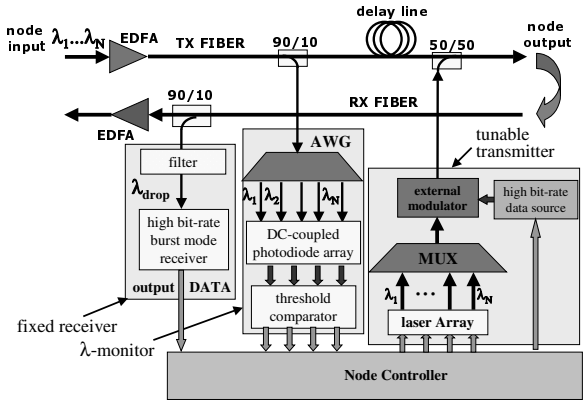

Fig. 2. OSATE switch port structure

To avoid the Head of the Line (HoL) [4] problem, each port 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.

Although in general $W \leq N$, we restrict our attention in this paper to the case $N = W$, in which each wavelength channel is dedicated to bring information to one switch output. In this case the architecture is fully non-blocking (like a crossbar): in each time slot, each input can deliver a packet to a different output, i.e., an input/output permutation can be served. However, due to the folded bus topology, the input/output permutation is sequentially built by successive decisions at ports according to the physical position along the transmission ring, while in traditional electronic crossbars packets for a given time slot are transmitted in parallel.

A collision may arise when an input tries to insert a packet on an already used time slot. Access decisions are based on channel inspection capability (similar to the carrier sense functionality in Ethernet), called λ-monitor. Thus, each input port knows which wavelengths have not been used by upstream inputs during the current time slot, and priority is given to in-transit traffic, i.e., a *multi-channel empty-slot* protocol is used.

While the sequential access decision permit the implementation of distributed control schemes, using a simple empty slot scheme might lead to fairness problems due to different access opportunities 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 [5], 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 scheduling algorithm must be able to ensure high throughput, bounded delays and equal transmission opportunity even when inputs are heavily loaded.

Although distributed algorithms are usually simpler to implement, requiring little or no control information exchange, they might show limited performance. On the contrary, centralized schemes are usually more complex, require a larger

information exchange, and may increase latencies, but can easily achieve high throughput. In the following, we describe the distributed Multi-Fasnet access scheme and compare its performance to those of well-known centralized scheduling algorithms, like iSLIP [6] and the throughput optimal Maximum Weight Matching [4].

III. The Fasnet Protocol

Fasnet [7] is an access protocol originally designed to guarantee fairness on a slotted dual bus topology. First, we adapt the protocol to a single channel folded bus topology; next, we extend it to a multichannel architecture.

Fasnet is an implicit token passing protocol developed to efficiently use channel capacity, providing fairness in resource sharing. To implement Fasnet, all ports have to listen on the transmission channel, excluding the first port in the transmission bus, dubbed master, which has to listen on the reception bus. As shown in Fig. 2, all ports are equipped with a λ-monitor capability to sense the transmission channel. However, the master monitoring function can be easily implemented by attaching, possibly with an optical switch, the λ-monitor to the reception bus. 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, and by all packets transmitted by switch ports after the locomotive. The master 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 port 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 port 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 port 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.

Fasnet is not able to reach 100% throughput, due to the idle time between two successive cycles. Indeed, the master recognizes the end of the train only when the last transmitted packet is sensed on the reception channel; this implies that a new locomotive is sent when no packets are traveling on the bus. Thus, the maximum achievable throughput, in overload, 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 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. Thus, the maximum achievable throughput under uniform traffic is given by:

$$TH_{max} = \frac{N \times Q}{N \times Q + \lceil 2 \times RTT \rceil + 1} \tag{1}$$

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

If we assume that the fabric is not overloaded, i.e., each port empties its queue without exhausting its quota, we can easily estimate the worst case access delay. This happens when a packet arrives immediately after the channel release; the port has to wait for the next train to transmit. Therefore, the worst case access delay at low loads can be the evaluated as:

$$D_{WC} \approx N \times Q^* + \lceil 2 \times RTT \rceil + 1 \tag{2}$$

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

$$Q^* = \frac{\rho}{1-\rho} \times \frac{\lceil 2 \times RTT \rceil + 1}{N} \tag{3}$$

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

In summary, Fasnet performance is limited both in throughput and in delay by the channel idle time needed by the master to detect the end of the current cycle. However, because of switching fabric dimensions, this idle time is of the order of few μs in a switch; thus, this problem is not a major issue in the studied scenario.

A. Multi-Fasnet Protocol

In a multichannel scenario, Fasnet behavior is replicated over the different wavelengths and W trains, one for each channel, are traveling on each bus. In the same time slot many channels might become available all together and a *train collision* happens. However, ports are equipped with only one tunable transmitter (see Fig. 2); thus, only one packet can be transmitter per time slot. Ports select the longest queue among the one associated with the available channels for transmission.

This means that ports may release a channel although they still have both quota and packets to transmit because of train collision; this is a peculiar behavior of multichannel environments. To guarantee throughput and fairness, ports are allowed to cumulate the unused quota: on the next cycle, at most Q packets plus the remaining quota of the previous cycle can be transmitted. To avoid excessive quota accumulation, the maximum quota that can be accumulated on a channel is bounded by either the current queue length on the corresponding channel, or by $M \times Q$, where M is a parameter set to 5 in simulation experiments (after some tuning).

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

$$TH_{max} = \frac{1}{W} \times \sum_{w=1}^{W} \frac{\sum_{i=1}^{N} \lambda_{iw} \times Q}{\sum_{i=1}^{N} \lambda_{iw} \times Q + \lceil 2 \times RTT \rceil + 1} \tag{4}$$

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

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

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

where Q^*_{iw} is the effective average quota used by port i on channel w.

Therefore, also in a multichannel scenario, the Multi-Fasnet performance is limited by the channel idle time.

IV. Performance Evaluation

We present performance results obtained by simulation considering a switching fabric with $W = 16$ wavelengths and $N = 16$ ports. The inter-port distance is about 100 ns and each port introduces a delay of 100 ns to perform the void detection; thus, the RTT of each ring is equal to 3.1 μs. Each port keeps W separate FIFO queues, one for each channel, with a queue size of about 32000, fixed size, packets.

We compare Multi-Fasnet access strategy with classic scheduling algorithms like iSLIP, as a representative of the class of heuristic but implementable scheduling algorithms, and the throughput optimal Maximum Weight Matching (MWM) scheduler.

Both uniform traffic and unbalanced traffic scenarios are considered. To describe the traffic scenarios, let ρ_i be 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 case, each input port transmits with probability $\lambda_{ij} = \rho_i \times 1/N$ to each output port. 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:

$$\lambda_{ij} = \begin{cases} \rho_i \times \frac{2}{3} & \text{if } j = i \\ \rho_i \times \frac{1}{3} & \text{if } j = |i+1|_N \\ 0 & \text{otherwise} \end{cases} \qquad (6)$$

where $|x|_N = x \bmod N$ (remainder of the division by N). In other words, port i only has traffic for output port i and $|i+1|_N$. For the log-diagonal traffic scenario, $\lambda_{ij} = 2 \times \lambda_{i|i+1|_N}$ and $\sum_j \lambda_{ij} = \rho_i$. In this case, the traffic is logarithmically skewed but the traffic is directed to all output ports.

We mainly focus on delay vs. throughput plots, obtained by simulation. Simulation runs exploit a proprietary simulation environment developed in the C language. Statistical significance of the results is assessed by running experiments with an accuracy of 3% under a confidence interval of 95%.

Fig. 3 shows the performance of Multi-Fasnet, iSLIP and MWM algorithms under the uniform traffic pattern. MWM and iSLIP achieve 100% throughput, while Multi-Fasnet performance are strongly affected by the value of the quota. As discussed in Sec. III-A, the larger the quota the larger the fabric utilization, since the idle time between the two consecutive train cycles has a lower impact as the train length increases. The maximum achievable throughput evaluated using (4) is,

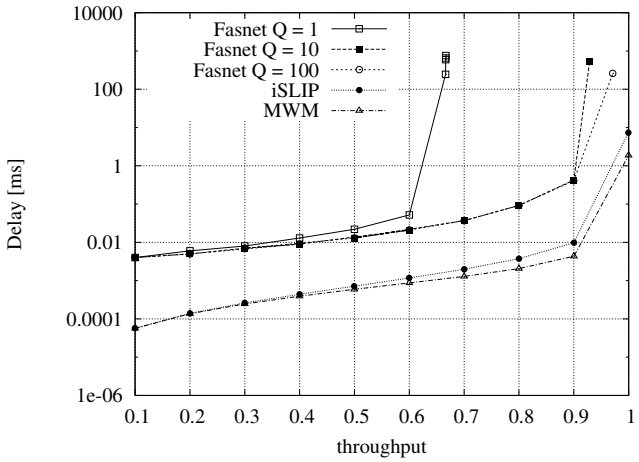

Fig. 3. Multi-Fasnet, iSLIP and MWM performance under uniform traffic

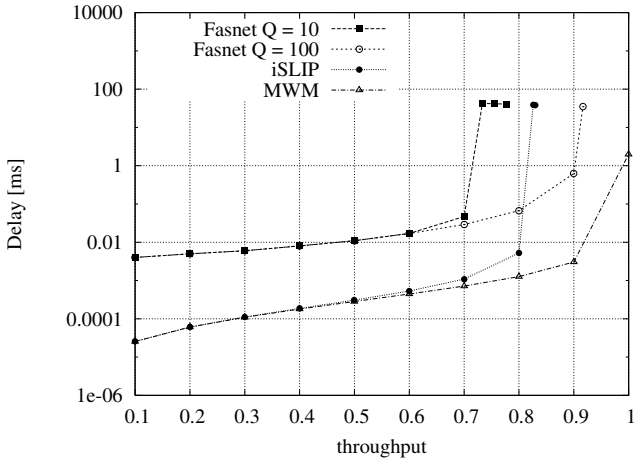

Fig. 4. Multi-Fasnet, iSLIP and MWM performance under log-diagonal traffic

respectively, $TH_{max} = 0.67$ for $Q = 1$, $TH_{max} = 0.95$ for $Q = 100$ and $TH_{max} = 0.995$ for $Q = 100$; these values are very close to those obtained by simulation: the minor differences are mainly due to the train collision effect.

When the fabric is lightly loaded, the average transmission delay is independent of the quota value; indeed, the train length depends on input traffic and fabric dimension only. With respect to a centralized scheme, the Multi-Fasnet protocol shows a larger transmission delay, equal to $2 \times RTT$ μs slots, which matches the idle time between two cycles, as explained in Sec. III. In overload 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 ports access the channel after $D_{WC_k} = N \times Q + \lceil 2 \times RTT \rceil + 1$ μs (slots) and transmit Q packets: the mean delay is equal to $D_{WC_k}/Q \times QL$ μs. Thus, the mean delay in overload conditions is approximately equal to 768 ms for $Q = 1$, 537 ms for $Q = 10$ and 514 ms for $Q = 100$.

Let compare now Multi-Fasnet with iSLIP and MWM under the two unbalanced traffic scenarios. Fig. 4 and Fig. 5 show the

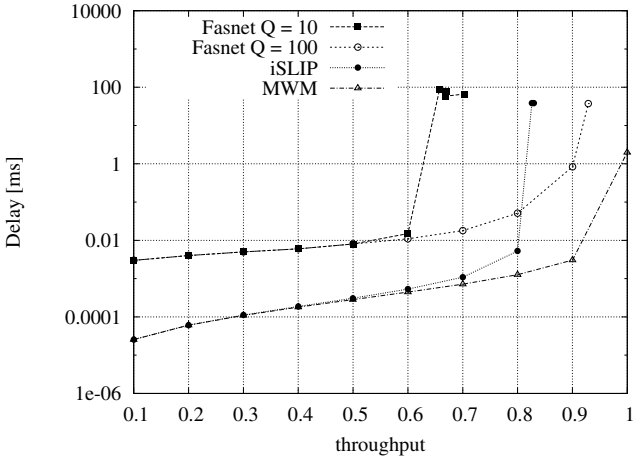

Fig. 5. Multi-Fasnet, iSLIP and MWM performance under bi-diagonal traffic

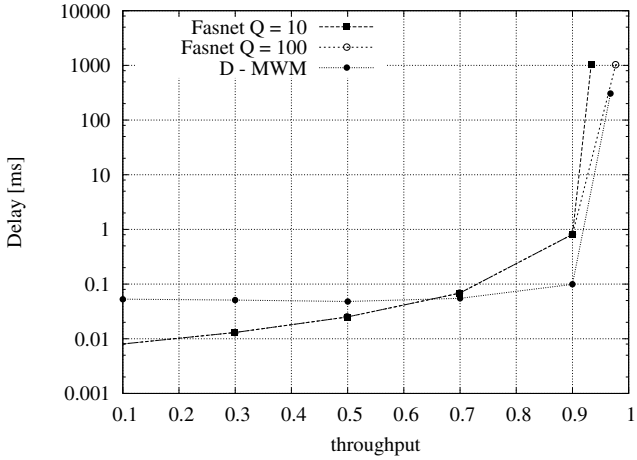

Fig. 6. Multi-Fasnet and MWM performance under uniform traffic when considering a real implementation of a centralized scheduler

delay vs. throughput plot under the bi-diagonal and the log-diagonal traffic scenarios, respectively. Although the Multi-Fasnet protocol always shows a larger delay when the fabric is lightly loaded due to the idle time between two trains, as the load increases, the differences between iSLIP and Multi-Fasnet decrease, and with a quota large enough, Multi-Fasnet is able to achieve a larger throughput than iSLIP.

Whereas Multi-Fasnet performance are obtained when running the scheduler over the proposed architecture, the considered centralized schedulers run under an idealized scheme: no delay is computed to transfer the information among ports and the centralized scheduler. In other words, the scheduler has an instantaneous view of all queues. Moreover, signaling messages needed to exchange information are not considered. However, in a real architecture, moving information from switch ports to the centralized scheduler, which normally runs in a dedicated card, implies a signaling delay. Furthermore, signaling messages require either a dedicated channel and an additional transceiver in each port, or must share the bandwidth with the data traffic. Thus, the comparison is rather biased in favor of the centralized schemes.

To have a fairer comparison, we defined a possible implementation of a centralized scheme in our architecture. We assume that the centralized scheduler is located at the head of the transmission ring. For simplicity, we rely on a dedicated channel to transfer the signaling information both from the scheduler to data ports (in a broadcast manner), and from data ports to the scheduler. As already noted, having extra transmission resources for signaling purposes implies additional complexity with respect to distributed schemes such as Multi-Fasnet. Each node signals its queue status to the scheduler every $1/N$ slots.

Performance results of the MWM scheduler when considering the above described implementation are shown in Fig. 6. Multi-Fasnet shows now quite remarkably performance, very close to the one obtained when running the throughput optimal MWM scheduler. Similar results, not shown for the sake of brevity, hold when considering the unbalanced traffic scenarios.

V. Conclusions and Future Work

We introduced a particular WDM, ring-based, distributed switching fabric called OSATE. We discussed the Multi-Fasnet strategy, obtained adapting to the WDM scenario an existing fairness control protocol, and we compared its performance against those of well-known centralized schedulers.

Simulation results show how Multi-Fasnet performance are slightly limited by the channel idle times between two consecutive cycles; thus, Multi-Fasnet needs large quota to reach large throughput. Despite these limitations, Multi-Fasnet achieves large utilization, it is very simple to implement in the proposed optical distributed switch and permits a distributed implementation, scaling better than the currently used centralized schemes.

Acknowledgment

The work described in this paper was performed 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.

References

[1] A.Bianco, D.Cuda, J.M.Finochietto, F.Neri, C.Piglione, "Multi-Fasnet Protocol: Short-Term Fairness Control in WDM Slotted MANs," *IEEE GLOBECOM 2006*, San Francisco, CA, USA, 27-30 November 2006.

[2] The OSATE Project: http://www.tlc-networks.polito.it/projects/osate/

[3] A.Carena, V.De Feo, J.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 (JSAC),* Vol. 22, No. 8, pp. 1561-1571, Oct. 2004

[4] 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.

[5] 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.

[6] N.McKeown,"The iSLIP Scheduling Algorithm for Input-Queued Switches", *IEEE/ACM Trans. on Networking* , Vol. 7, No. 2, April 1999.

[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.