An Optical Interconnection Architecture for Large Packet Switches

A. Bianco, E. Carta, D. Cuda, J. M. Finochietto, and F. Neri
Dipartimento di Elettronica, Politecnico di Torino, 10129 Torino, Italy

ABSTRACT

The design of switching architectures for today's telecommunication networks needs to consider the limits imposed by electronic technology; in particular, it must take into account power consumption and dissipation, as well as power supply and footprint requirements. Currently, optical technology is exploited mainly for transmission over optical links requiring large bandwidth-distance products; however, many researchers and switch architects believe that its introduction for switching functions can overcome most of the current design limits. Since many years, the research community has been studying not only solutions that make use of optics inside electronic switches but that also switching architectures that implement optical switching without any need of optoelectronic conversion. In this paper, we propose an optical switching architecture based on a WDM (Wavelength Division Multiplexing) optical bus, aiming at interconnecting electronic line cards inside packet switches. In particular, the use of distributed packet scheduling techniques is addressed and its performance is discussed.

Keywords: switching architectures, optical switching, packet scheduling.

1. INTRODUCTION

Although routers with capacities above several Tb/s are currently available, the number of backplane interconnections and the power density are reaching physical limits. Indeed, each new generation of router consumes more power than the previous one, and it is more difficult to package a router in one single rack of equipment. Thus, high-end routers often comprise several racks of equipment: one or more racks host the electronic switching fabric and the control logic, while others racks, the line cards. In this configuration, optical links start being used to interconnect the fabric and line cards. These solutions occupy valuable space, consume too much power, and pose reliability concerns due to the large number of active components in the switch fabric.

Routers with optical interconnection architectures, can scale better to higher capacities, increase reliability, and at the same time significantly reduce footprint and power consumption [1]. Moreover, since current routers make use of centralized electronic arbitration schemes, their performance can be limited by the complexity of implementing these arbitration schemes as the aggregate packet processing rate increases [2]. As a result, the value of an optical interconnection architecture is closely related to the availability and performance of distributed arbitration schemes.

In this paper we proposed an optical interconnection architecture for large packet switches, originally conceived, studied and prototyped as a network architecture for the metro area [3]. However, as it will be discussed in the next sections, the foundations of this architecture can be considered and extended for building scalable optical interconnects where distributed scheduling algorithms can be implemented. The paper is organized as follows. Section 2 introduces the proposed architecture and describes the use of a distributed scheduling algorithm. Section 3 discusses simulation results of the performance of distributed and centralized scheduling schemes. Section 4 analyzes the scalability of the proposed architecture considering physical issues and proposes possible improvements to the original architecture. Finally, Section 5 concludes the paper.

2. SYSTEM ARCHITECTURE

The proposed architecture is based upon a passive WDM (Wavelength Division Multiplexing) all-optical data path over a folded bus as depicted in Fig. 1*a*. The folded bus conveys W wavelengths which first traverse the transmission (TX) bus and, after a folding point, the reception (RX) bus. Each of the N line cards attached to the bus is equipped with one transmitter and one receiver operating at the data rate of a single WDM channel. The input and output ports of the line card are passively coupled to the TX bus and RX bus respectively. Since full connectivity between all available line cards must be provided on a packet-by-packet basis, fast wavelength tunability at transceivers is required to temporally allocate all-optical single-hop bandwidth. However, due to cost tunability of transceivers, this is limited only to transmitters, while receivers are permanently tuned to a specific WDM channel. When a single receiver per WDM channel is present, and thus the number of available WDM channels W equals the number of line cards N, the architecture can be shown to be equivalent to a distributed crossbar switch, which is able to connect at every time up to N disjoint input-output pairs.

The architecture is synchronous, with time slotted operation. For this purpose, one additional wavelength of the WDM comb is dedicated to the distribution of synchronization information along the data path from the first

The work reported in this paper was supported by the European Network of Excellence e-Photon/ONe+.

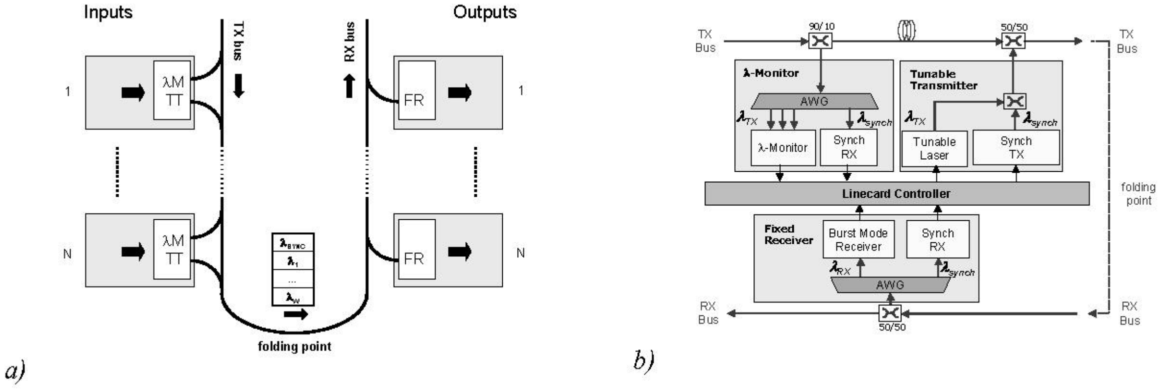

Figure 1. a) Interconnection architecture b) Line card architecture.

input port to the last output port. As a result, each line card. as shown in Fig. 1*b*, is equipped with one tunable transmitter (TT) which can also include the synchronization transmitter, and one fixed receiver (FR) operating as a burst mode receiver, which drops the synchronization signal to facilitate clock and data recovery. The slotted behavior also facilitates the use of distributed scheduling mechanisms, which avoid packet collisions by means of a void-detection mechanism or λ-Monitor (λM) by which line cards know which wavelengths were not used by upstream line cards in each time slot. Priority is given to in-transit traffic, thus without requiring any packet buffering for in-transit data, as well as no packet switching or stripping in the optical domain. The TT is used to insert packets on free time slots on the wavelength leading to the packet's destination. The FR receives all packets on its assigned wavelength. Since output always receive packets on the same wavelength, thus in a non-overlapping way, receiver contention does not occur, since it is solved at the transmitter side. Head-of-the-Line (HoL) blocking is avoided since Virtual Output Queuing (VOQ) is envisioned at inputs; thus, line cards queue packets to be inserted on a per destination-wavelength basis. Since neither packet collisions nor receiver contention occur, transmitted packets are never lost except for transmission errors in the physical layer.

The folded bus topology imposes input ports to transmit packets sequentially and not in parallel like in traditional crossbars. However, a simple empty slot arbitration scheme might lead to fairness problems due to access probabilities depending on the position of the input ports along the TX bus. Referring to Fig. 1*a*, an upstream input can "flood" a given wavelength, reducing (or even blocking) the transmission opportunities of downstream ports competing for access to that channel, thus leading to significant fairness problems. Therefore, suitable scheduling algorithms are required to arbitrate packet transmissions to ensure not only high throughput and bounded delay, but also equal transmission probabilities for all input ports, even when some inputs are heavily loaded. These arbitration schemes can be either centralized or distributed. The former requires the use of an electronic scheduler that, after receiving status information from line cards, decides a new permutation, i.e., input/output port connection pattern, for each time slot. The latter uses only locally available information on line cards to determine which packet to transfer. Thus, centralized schemes can potentially offer better performance in terms of throughput and latency than distributed ones, but electronic complexity of the scheduler implementation must also be taken into account. Indeed, optimal algorithms such as Maximum Weight Matching (MWM) [4] are impractical because of their complexity, and sub-optimal ones, such as iSLIP [5], are in general preferred.

In this context, the performance of a distributed scheduling scheme becomes a crucial issue to assess the actual efficiency of the proposed optical interconnection architecture. One of these schemes is dubbed Fasnet [6], and it was previously proposed and studied to arbitrate access in our architecture [7]. As a first approximation, the Fasnet scheme behaves as a distributed polling mechanism without the need of a centralized scheduler. Fasnet operates cyclically, and each cycle is associated with a chained transmission of packets by all line cards, named *train*. A train is composed by a first control packet, dubbed locomotive, transmitted by the first line card, and by all packets transmitted sequentially by line cards after the locomotive. This first line card starts a new cycle, thus transmits a new locomotive, every time it detects the end of a returning train (i.e., an empty slot on the RX bus). To limit the maximum number of packets each line card can transmit on each cycle, each line card is assigned a quota Q. When a line card senses an end of train, i.e., an empty slot on the TX bus, 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 for that channel. Once a line card 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.

In a WDM multi-channel network, the Fasnet behavior should be replicated over all available wavelengths: thus, W trains exist, one for each channel. Note that the Fasnet scheme permits the transmission of variable-size packets despite the slotted behavior of the network. In fact, variable-size packets can easily fit in successive

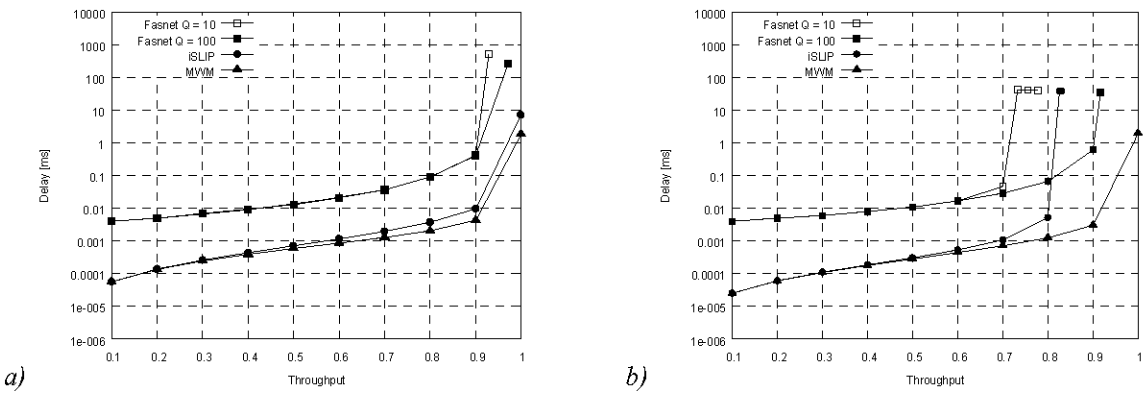

Figure 2. Performance under a) uniform traffic b) log-diagonal traffic.

time-slots without collisions problems, since each input card has full control of the channel access during transmission. However, Fasnet is not able to reach 100% throughput, due to the idle time between two successive cycles. Indeed, the first line card detects 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 only transmitted when no more packets are traversing the folded bus. Thus, the maximum achievable throughput, in highly loaded traffic conditions, is affected by the ratio between the maximum train length, equal to $N \times Q$, and the cycle duration, which is equal to $N \times Q$ plus the time needed by the first line card to detect the end of the current train. In our architecture, this starvation time is the folded bus propagation delay. As a result, larger values of Q increase the maximum achievable throughput but may impose larger packet transfer delays as line cards may have to wait for longer periods of time to have a chance to schedule new arriving packets. Besides the value of Q, note that, in lightly loaded traffic conditions, a newly arrived packet may not be scheduled immediately (although no other packet is being transmitted) since line cards have to wait in the worst case one starvation time to schedule packets.

3. PERFORMANCE EVALUATION

We evaluate the throughput and delay performance of Fasnet in the proposed architecture. Since performance depends on the value of quota Q, different Q values have been considered. Besides, we compare Fasnet with classical centralized scheduling schemes like iSLIP and MWM. These results have been obtained by simulation considering an architecture with 16 line cards (N = W = 16). The distance between line cards is 20 meters (100 ns) and each line card introduces a delay of 100 ns to perform the void detection function; thus, the folded bus propagation time is about 7 μs. We assume fixed-size packets that fit in one time slot; the time slot duration is 1 μs. Each of the N separate FIFO queues on each line card can store up to 32 K packets. Two traffic scenarios are considered: uniform traffic and log-diagonal traffic. To describe the traffic scenarios, let p_{ij} be the probability that a packet arriving at input port i is addressed to output port j. In the uniform traffic case, the whole network capacity is equally shared by all line cards, i.e., each input port transmits with probability $p_{ij} = 1/N$ to all other ports. In the log-diagonal traffic pattern, $p_{ij} = 2 \times p_{ii'}$ where $i' = |i+1|_N$.

Performance under the uniform traffic scenario is shown in Fig. 2*a*. At a first glance, centralized schemes offer better performance; however, the Fasnet scheme with Q = 100 is able to reach a throughput quite close to that of centralized schemes, while for a smaller quota value (Q = 10) the performance is limited to about 90% of the capacity. The main difference between centralized schemes and Fasnet is related to delays. Indeed, even at very low loads the Fasnet scheme imposes a packet delay of about 7 μs due to the starvation time between cycles. In Fig. 2*b* we show results for the log-diagonal case. As in the uniform scenario, packets experience larger delays when scheduled under the Fasnet scheme rather than with centralized ones. Although the optimal MWM scheme is still able to offer 100% throughput, the iSLIP only achieves about 80% due to its sub-optimal nature. On the contrary, the Fasnet scheme with Q = 100 outperforms iSLIP in terms of throughput achieving almost 90%, a quite remarkable result. Note that Fasnet is fully distributed, and no state information is to be distributed. Instead, both centralized schemes require to exchange line card state information, which translates into the need for an additional signaling channel and signaling transceivers. According to the above discussed results, the extra costs of centralized schemes do not seem to pay off in terms of performance.

4. SCALABILITY ANALYSIS

Although the folded bus architecture shown in Fig. 1*a* is a promising architecture to implement distributed scheduling algorithms, it presents physical scalability limitations in terms of the number of ports N that can be supported, due to the sequential coupling of line cards. As illustrated in Fig. 1*b*, each line cards introduces at

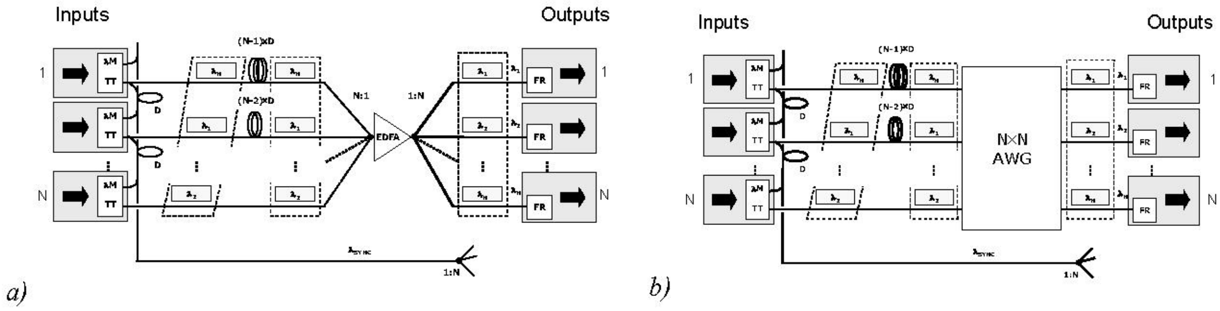

Figure 3. Interconnection architecture a) with amplifiers b) with AWG.

least 6 dB of power loss due to the coupling of its transceivers to the folded bus. Thus, the scalability is strongly limited by the total power loss that is equal to *6 dB* $\times N$. If EDFA optical amplifiers are cascaded to overcome power losses, then the scalability of the architecture becomes limited by ASE noise accumulation. As a result, in this section we discuss possible evolutions of the proposed architecture that can overcome these scalability limitations while keeping the low complexity of the original architecture. In particular, we aim at designing more scalable architectures where the Fasnet scheme could still be used to schedule packets.

In Fig. 3 we sketch two possible evolutions of the interconnection architecture. In both cases, a major change is the space separation of the data path channel from the synchronization one, transported in different fibers. Packet scheduling is ruled by the synchronization signal that gets delayed by a time D between each line card. As a result, line cards schedule packets at different times, but due to proper delays inserted in the data path these packets get aligned on the same time slot. Since line cards need to monitor the current wavelengths (i.e., empty/free), the data path is passively coupled to the synchronization fiber in order to insert a sample of the transmitted power in the data path. In this way, the wavelengths' state information can be read by the λ-Monitor on each time slot from the synchronization fiber. Although line cards have the same complexity and behavior of the original architecture, the data path can be designed in a completely different way. Indeed, instead of a folded bus physical topology a tree one can be considered as shown in Fig. 3*a*. In this way, packets experience a nominal power loss of 6 dB $\times \log_2 N$ instead of 6 dB $\times N$ as in the original architecture. Besides, for large N, this loss can be compensated by one or few EDFA optical amplifiers. An alternative to this architecture is shown in Fig. 3*b* where a $N \times N$ AWG is introduced to route each wavelength to an output port. Thus, packets experience the AWG insertion loss which typically is between 2 – 8 dB (instead of the 6 dB $\times \log_2 N$ of the two couplers). In conclusion, these architectures can overcome power budget limitations of the original one, and offer scalability in terms of the number of line cards N equal to the maximum number of wavelengths W that can be supported.

5. CONCLUSIONS

In this paper we have proposed an optical architecture to interconnect electrical line cards in large packet switches. As a major value of the proposed architecture, a distributed arbitration scheme has been proposed that offers low implementation complexity, low control latency and high scalability. Besides, its performance has been shown to be comparable with that of centralized arbitration schemes. Finally, different architectural variants of the proposed architecture have been introduced to improve scalability issues.

REFERENCES

[1] J. Gripp, *et al.*; Optical switch fabrics for ultra-high-capacity IP routers, *J. Lightwave Technol.*, vol. 21, no.11, pp. 2839- 2850, Nov. 2003.

[2] N. McKeown: Optics inside Routers, in *Proc. ECOC 2003*, Rimini, Italy, Sep. 2003.

[3] A. Carena, *et al.*; RingO: an experimental WDM optical packet network for metro applications, *Selected Areas in Communications, IEEE Journal on*, vol.22, no.8, pp. 1561-1571, Oct. 2004.

[4] N. McKeown, *et al.*; Achieving 100% throughput in an input-queued switch, *IEEE Transactions on Communications*, vol. 47, no. 8, pp. 1260-1267, Aug. 1999.

[5] N. McKeown, The iSLIP Scheduling Algorithm for Input-Queued Switches", *IEEE/ACM Transactions on Networking*, vol. 7, no. 2, April 1999.

[6] J. O. Limb, *et al.;* Description of Fasnet – A Unidirectional Local–Area Communication Network, *The Bell System Technical Journal*, vol. 61, no. 7, September 1982.

[7] A. Bianco, *et al.*; Multi-Fasnet Protocol: Short-Term Fairness Control in WDM Slotted MANs, *Global Telecommunications Conference, 2006. GLOBECOM '06. IEEE*, pp. 1-5, Nov. 2006.