Part I

Enabling technologies

1 Optical switching fabrics for terabit packet switches

Davide Cuda, Roberto Gaudino, Guido A. Gavilanes Castillo, and Fabio Neri

Politecnico di Torino, Turin, Italy

A key element of past, current, and future telecommunication infrastructures is the switching node. In recent years, packet switching has taken a dominant role over circuit switching, so that current switching nodes are often packet switches and routers. While a deeper penetration of optical technologies in the switching realm will most likely reintroduce forms of circuit switching, which are more suited to realizations in the optical domain, and optical cross-connects [1, Section 7.4] may end up playing an important role in networking in the long term, we focus in this chapter on high-performance packet switches.

Despite several ups and downs in the telecom market, the amount of information to be transported by networks has been constantly increasing with time. Both the success of new applications and of the peer-to-peer paradigm, and the availability of large access bandwidths (few Mb/s on xDSLs and broadband wireless, but often up to 10's or 100's of Mb/s per residential connection, as currently offered in Passive Optical Networks – PONs), are causing a constant increase of the traffic offered to the Internet and to networking infrastructures in general. The traffic increase rate is fast, and several studies show that it is even faster than the growth rate of electronic technologies (typically embodied by Moore's law, predicting a two-fold performance and capacity increase every 18 months).

Optical fibers are the dominant technology on links between switching nodes, and, while the theoretical capacity of several tens of Tb/s on each fiber is practically never reached, still very high information densities are commercially available: 10–40 Gb/s, and soon 100 Gb/s, per wavelength channel are commonplace in WDM (Wavelength Division Multiplexing) transmission systems capable of carrying up to tens of channels on a single fiber, leading to Tb/s information rates on a single optical fiber.

Very good and mature commercial offerings are on the market today for packet switches and routers, with total switching capacities up to few Tb/s. These devices are today fully realized in the electronic domain: information received from optical fibers is converted in linecards to the electronic domain, in

Next-Generation Internet Architectures and Protocols, ed. Byrav Ramamurthy, George Rouskas, and Krishna M. Sivalingam. Published by Cambridge University Press. © Cambridge University Press 2011.

which packets are processed, stored to solve contentions, and switched through a switching fabric to the proper output port in a linecard, where they are converted back to the optical domain for transmission.

The fast traffic growth however raises concerns on the capability of electronic realizations of packet switches and routers to keep up with the amount of information to be processed and switched. Indeed, the continuous evolution of high-capacity packet switches and routers is today bringing recent realizations close to the fundamental physical limits of electronic devices, mostly in terms of maximum clock rate, maximum number of gates inside a single silicon core, power density, and power dissipation (typically current large routers need tens of kW of power supply; a frequently cited example is the CRS-1 System – see [2, page 10]). Each new generation of switching devices shows increasing component complexity and needs to dissipate more power than the previous one. The current architectural trend is to separate the switching fabric from the linecards and often to employ optical point-to-point interconnections among them (see [3, page 6] again for the CRS-1). This solution results in a large footprint (current large switches are often multi-rack), poses serious reliability issues because of the large number of active devices, and is extremely power-hungry.

A lively debate on how to overcome these limits is ongoing in the research community. Optical technologies today are in practice limited to the implementation of transmission functions, and very few applications of photonics in switching can be found in commercial offerings. Several researchers however claim that optical technologies can bring significant advantages also in the realization of switching functions: better scalability towards higher capacities, increased reliability, higher information densities on internal switch interconnections and backplanes, reduced footprint and better scaling of power consumption [7, 8].

In this chapter[1] we consider photonic technologies to realize subsystems inside packet switches and routers, as recently has been done by several academic and industrial research groups. In particular, we refer to a medium-term scenario in which packet switching according to the Internet networking paradigm dominates. Hence we assume that packets are received at input ports according to current formats and protocols (such as IP, Ethernet, packet over Sonet). We further assume that packets are converted in the electronic domain at linecards for processing and contention resolution. We propose to use optical interconnections among linecards, hence to implement an optical switching fabric internally to the switch. At output linecards, packets are converted back to legacy formats and protocols, so that the considered architectures remain fully compatible with current network infrastructures. For these architectures we will evaluate the maximum achievable switching capacities, and we will estimate the costs (and their scaling laws) of implementations based on currently available discrete components.

[1] Preliminary parts of this chapter appeared in [4, 5, 6]. This work was partially supported by by the Network of Excellence BONE ("Building the Future Optical Network in Europe"), funded by the European Commission through the Seventh Framework Programme.

1.1 Optical switching fabrics

To study the suitability of optical technologies for the realization of switching fabrics inside packet switching devices, we focus on three optical interconnection architectures belonging to the well-known family usually referred to as "tunable transmitter, fixed receiver" (TTx-FRx), which was widely investigated in the past (the three specific architectures were studied [9, 10] within the e-Photon/ONe European project). In particular, we consider optical interconnection architectures based on the use of broadcast-and-select or wavelength-routing techniques to implement packet switching through a fully optical system in which both wavelength division and space multiplexing are used. Indeed, WDM and space multiplexing have proven to make best use of the distinctive features of the optical domain.

Our architectures are conceived in such a way that switching decisions can be completely handled by linecards, enabling the use of distributed scheduling algorithms. We will, however, not deal with control and resource allocation algorithms, thus focusing on the switching fabric design, for which we will consider in some detail the physical-layer feasibility, the scalability issues, and the costs related to realizations with currently available components.

Many optical switching experiments published in the last 10–15 years have used optical processing techniques such as wavelength conversion or 3R regeneration [1, Chapter 3], and even optical label recognition and swapping, or all-optical switch control, and have been often successfully demonstrated in a laboratory environment. However, they are far from being commercially feasible, since they require optical components which are still either in their infancy or simply too expensive. We take a more conservative approach, restricting our attention to architectures that are feasible today, as they require only optical components commercially available at the time of this publication. Fast-tunable lasers with nanosecond switching times are probably the only significant exception, since they have not yet a real commercial maturity, even though their feasibility has already been demonstrated in many experimental projects [7] and the first products are appearing on the market.

The reference architecture of the optical switching fabric considered in this chapter is shown in Figure 1.1: a set of N input linecards send packets to an optical interconnection structure that provides connectivity towards the N output linecards using both WDM and optical space multiplexing techniques. The optical switching fabric is organized in S switching planes, to which a subset of output linecards are connected. As we mainly refer to packet switching, fast optical switches (or in general plane distribution subsystems) allow input linecards to select the plane leading to the desired output linecard on a packet-per-packet basis. Obviously, slower switching speeds would suffice in case of circuit switching. Within each plane, wavelength routing techniques select the proper output. Thus packet switching is controlled at each input linecard by means of a fast tunable laser (i.e., in the wavelength domain) and, for $S > 1$, of a fast optical switch (i.e., in the space domain). Each linecard is equipped with one tunable transmitter (TTx) and one fixed-wavelength burst-mode receiver (BMR) operating at the data rate of a single WDM channel. Burst-mode operation is required on a packet-by-packet basis. Note that both TTx's and BMR's have recently appeared in the market to meet the demand for flexible WDM systems (for TTx's) and for upstream receivers in PONs (for BMR's).

For simplicity,[2] we assume that all the considered architectures have a synchronous and time-slotted behavior, as reported in [4] and [5]: all linecards are synchronized to a common clock signal which can be distributed either optically or electrically.

Packet transmissions are scheduled so that at most one packet is sent to each receiver on a time slot (i.e., contentions are solved at transmitters). Packet scheduling can be implemented in a centralized fashion as in most current packet switches. In this case, an electronic scheduler is required so that, after receiving status information from linecards, it decides a new permutation, i.e., an input/output port connection pattern, for each time slot. Centralized schemes can potentially offer excellent performance in terms of throughput, but the electronic complexity of the scheduler implementation can upper bound the achievable performance [8]. Furthermore, centralized arbitration schemes require signaling bandwidth to collect status information and to distribute scheduling decisions; these introduce latencies due to the time needed to propagate such information and to execute the scheduling algorithm. In this context, the implementation of a distributed scheduling scheme becomes a crucial issue to assess the actual value of proposed optical interconnection architectures. Distributed schemes that exhibit fair access among linecards using only locally available information (see, e.g., [11, 12]) have been proposed for architectures similar to those considered in this chapter; they avoid bandwidth waste due to the signaling process and limit the scheduling algorithm complexity, thus improving the overall fabric scalability.

The lasers tuning range, i.e, the number of wavelengths a transmitter is required to tune to, can be a practical limiting factor. Even for laboratory prototypes, the maximum tuning range for tunable lasers is in the order of a few tens of wavelengths [13]. As a result, the wavelength dimension alone could not ensure input/output connectivity when the number N of linecards is large. Multiple switching planes, i.e., the space diversity dimension, were indeed introduced to overcome this limitation. By doing so, since the same wavelengths can be reused on each switching plane, if S is the number of switching planes, a wavelength tunability equal to N/S (instead of N) is required.

In the three considered architectures, shown in Figs. 1.2–1.4, transmitters reach the S different planes by means of proper optical *distribution* stages that

[2] This assumption is not strictly necessary, but we introduce it to describe in a simpler way the operation of the switching fabric.

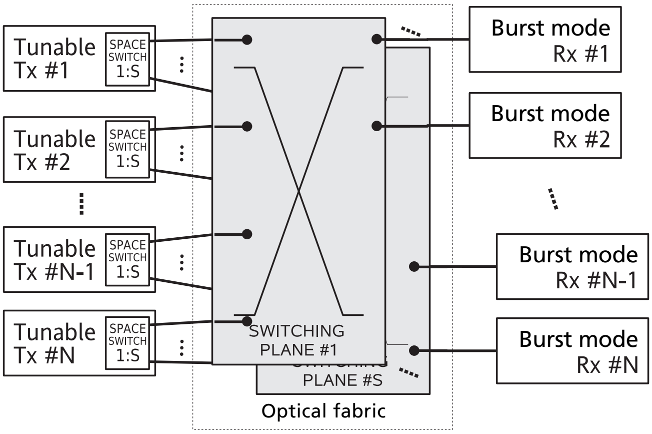

Figure 1.1 Multi-plane optical fabric architecture.

actually differentiate each of the three architectures. At the output of the switching plane, an $S:1$ optical coupler collects the packets, that are amplified with an Erbium Doped Fiber Amplifier (EDFA), and then distributed by a WDM demultiplexer to the N/S receivers. These fabric architectures are designed so that the number of couplers and other devices that packets have to cross is the same for all input/output paths. The EDFA WDM amplification stages can thus add equal gain to all wavelength channels since all packets arrive with the same power level.

1.1.1 Wavelength-selective (WS) architecture

This architecture, presented in Figure 1.2, was originally proposed for optical packet switched WDM networks inside the e-Photon/ONe project [9]. This optical switching fabric connects N input/output linecards by means of broadcast-and-select stages; groups of N/S TTx's are multiplexed in a WDM signal by means of an $N/S:1$ coupler; then this signal is split into S copies by means of a $1:S$ splitter. Each copy is interfaced to a different switching plane by means of wavelength selectors, composed by a demux/mux pair separated by an array of N/S Semiconductor Optical Amplifier (SOA) gates, which are responsible for both plane and wavelength selection.

Wavelength-selective architecture is a blocking architecture. Since within a group of input linecards all the transmitters must use a different wavelength, it is not possible for transmitters located in the same input group to transmit to output ports located on different switching planes that receive on the same wavelength. Proper scheduling strategies can partly cope with this issue, but we do not discuss them here.

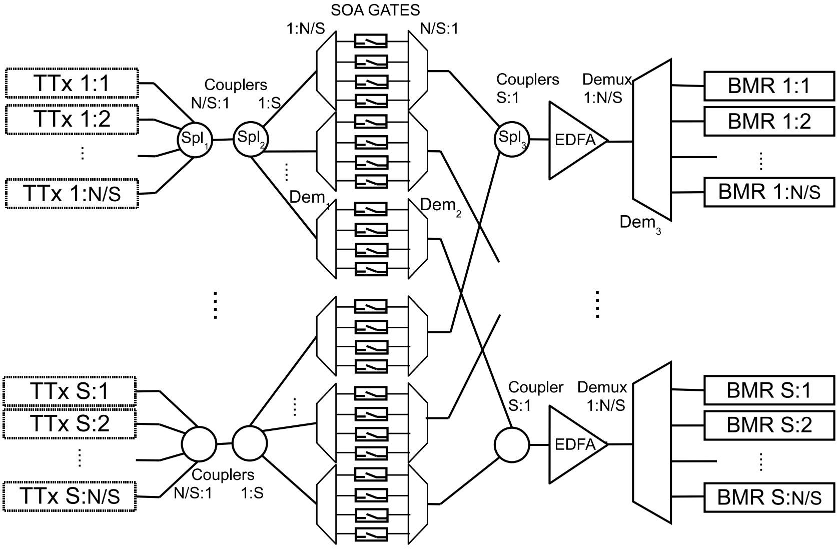

Figure 1.2 Wavelength-selective (WS) architecture.

1.1.2 Wavelength-routing (WR) architecture

In this case (see Figure 1.3) the wavelength-routing property of Arrayed Waveguide Gratings (AWGs), the main component of the distribution stage, is exploited to perform both plane and destination selection: no space switch is necessary. Each collecting stage gathers all the packets for a specific plane.

The transfer function of AWGs [14] exhibits a cyclic routing property: several homologous wavelengths belonging to periodically repeated Free Spectral Ranges (FSR) are identically routed to the same AWG output. The WR fabric can either exploit this property or not, with conflicting impacts on tuning ranges and crosstalk, respectively.

The former situation (i.e., full exploitation of the AWG cyclic property), called WR Zero-Crosstalk (WR-ZC) is depicted in Figure 1.3, which shows an instance of the WR architecture with $N = 9$ and $S = 3$. Note that each transmitter in a group of N/S transmitters uses N wavelengths in a different FSR; in other words we have N/S FSRs comprising N wavelengths each. The tunability range of each transmitter is N, and each receiver is associated with a set of N/S different wavelengths (for receivers this is not a real limitation because the optical bandwidth of photodiodes is actually very large). In this way, by exploiting the cyclic property of AWGs, we can prevent linecards from reusing the same wavelength at different input ports of the AWG. The advantage is that no coherent crosstalk (see [4]) is introduced in the optical fabric, and only out-of-band crosstalk is present, which introduces negligible penalty (see Section 1.2.1). Even though this solution can be almost optimal with respect to physical-layer impairments,

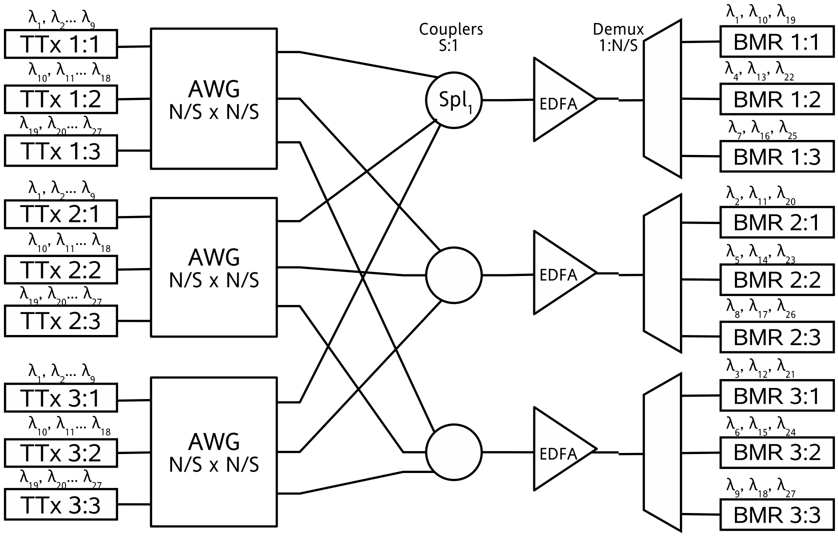
$\lambda_1, \lambda_2 \ldots \lambda_9$
TTx 1:1
$\lambda_{10}, \lambda_{11} \ldots \lambda_{18}$
TTx 1:2
$\lambda_{19}, \lambda_{20} \ldots \lambda_{27}$
TTx 1:3
$\lambda_1, \lambda_2 \ldots \lambda_9$
TTx 2:1
$\lambda_{10}, \lambda_{11} \ldots \lambda_{18}$
TTx 2:2
$\lambda_{19}, \lambda_{20} \ldots \lambda_{27}$
TTx 2:3
$\lambda_1, \lambda_2 \ldots \lambda_9$
TTx 3:1
$\lambda_{10}, \lambda_{11} \ldots \lambda_{18}$
TTx 3:2
$\lambda_{19}, \lambda_{20} \ldots \lambda_{27}$
TTx 3:3
AWG N/S x N/S
AWG N/S x N/S
AWG N/S x N/S
Couplers S:1
Spl_1
EDFA
EDFA
EDFA
Demux 1:N/S
$\lambda_1, \lambda_{10}, \lambda_{19}$
BMR 1:1
$\lambda_4, \lambda_{13}, \lambda_{22}$
BMR 1:2
$\lambda_7, \lambda_{16}, \lambda_{25}$
BMR 1:3
$\lambda_2, \lambda_{11}, \lambda_{20}$
BMR 2:1
$\lambda_5, \lambda_{14}, \lambda_{23}$
BMR 2:2
$\lambda_8, \lambda_{17}, \lambda_{26}$
BMR 2:3
$\lambda_3, \lambda_{12}, \lambda_{21}$
BMR 3:1
$\lambda_6, \lambda_{15}, \lambda_{24}$
BMR 3:2
$\lambda_9, \lambda_{18}, \lambda_{27}$
BMR 3:3

Figure 1.3 Wavelength-routing Zero-Crosstalk (WR-ZC) architecture.

the AWGs must exhibit an almost identical transfer function over N/S FSRs, and the EDFA amplifying bandwidth has to be significantly larger.

If instead the AWG cyclic property is exploited only in part (we call this architecture simply WR), and the system operation is limited to a single FSR (all TTx's are identical with tunability N), some in-band crosstalk can be introduced, which can severely limit scalability [4] when two or more TTx's use the same wavelength at the same time. However, proper packet scheduling algorithms, which avoid using the same wavelengths at too many different fabric inputs at the same time [15], can prevent the switching fabric from operating in high-crosstalk conditions.

1.1.3 Plane-switching (PS) architecture

In the PS fabric, depicted in Figure 1.4, the distribution stage is implemented in the linecards by splitting transmitted signals in S copies by a $1 : S$ splitter, and then sending the signals to SOA gates. A given input willing to transmit a packet to a given output must first select the destination plane by turning off all the SOA gates except the one associated with the destination plane and, then, use wavelength tunability to reach the desired output port in the destination plane. Coupling stages are organized in two vertical sections: the first is a distribution stage used by inputs to reach all planes, while the second is a collecting stage to gather packets for each destination plane. Each plane combines at most N/S packets coming from the N input linecards.

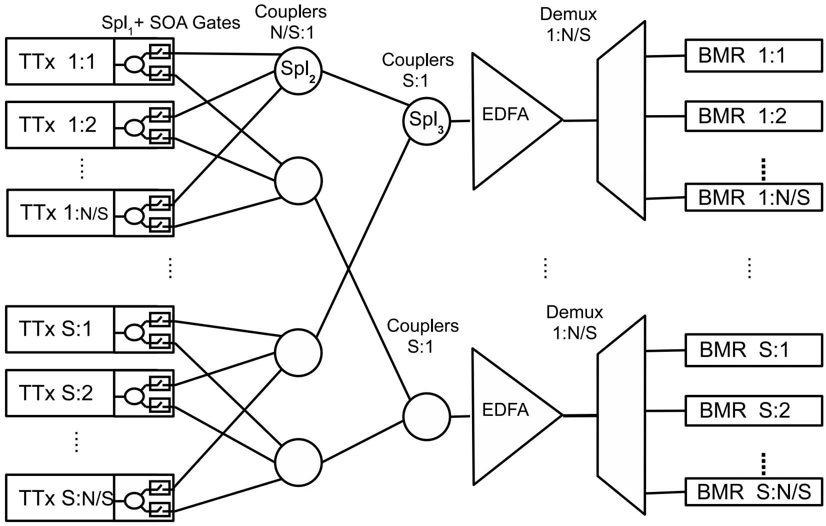

Figure 1.4 Plane-switching (PS) architecture.

1.2 Modeling optical devices

None of the proposed optical fabrics includes any signal regeneration besides pure linear optical amplification. Using the common terminology introduced in [1], we have at most 1R regeneration of the signals inside the optical fabric, while we exclude 2R and 3R regeneration. As a result, physical layer impairments may accumulate when increasing the port count N or the number of planes S, so that the characterization of the used optical devices becomes crucial to effectively assess each architecture's ultimate scalability. In performing the analysis described in this chapter, we observed that a first-order scalability assessment based on theoretical insertion loss values gives unrealistic results. As a clear example, the AWG in the WR architecture has an insertion loss that in a first approximation does not depend on the number of input/output ports, thus leading to a theoretical "infinite scalability." Clearly, we needed a more accurate second-order assessment capable of capturing other important effects that characterize commercial devices, such as polarization dependence, excess losses, channel uniformity, and crosstalk. Despite their different nature, all these effects can be expressed as an input/output equivalent power penalty which accounts for both actual physical power loss and the equivalent power penalty introduced by other second-order transmission impairments, as described below. We only focused our study on optical components, as fiber-related effects (e.g., dispersion, attenuation, non-linearities, cross-phase modulation, etc.) are likely to be negligible in the proposed architectures, mainly due to the short distances involved.

1.2.1 Physical model

The following physical-layer effects are taken into account in our analysis. See [4] for details.

Insertion Loss (IL): We indicate as insertion loss the total worst-case power loss, which includes all effects related to internal scattering due to the splitting process and also non-ideal splitting conditions, such as material defects, or manufacturing inaccuracies. In the case of n-port splitters, the splitting process gives a minimum theoretical loss increasing with $10 \log n$ dB, but extra loss contributions due to non-ideal effects, often referred to as Excess Losses (EL), must also be considered.

Uniformity (U): Due to the large wavelength range typically covered by multiport devices, different transmission coefficients exist for different wavelengths. Over the full WDM comb, the propagation conditions vary slightly from center channels to border ones. Similar uneven behaviors appear in different spatial sections of some components. These differences are taken into account by the U penalty component, which is often referred to as the maximum IL variation over the full wavelength range in all paths among inputs and outputs.

Polarization Dependent Loss (PDL): The attenuation of the light crossing a device depends on its polarization state due to construction geometries, or to material irregularities. Losses due to polarization effects are counted as a penalty in the worst propagation case.

Crosstalk (X): A signal out of a WDM demultiplexing port always contains an amount of power, other than the useful one, belonging to other channels passing through the device. This effect is generally referred to as crosstalk. For a given useful signal at wavelength λ, the crosstalk is usually classified [1] as either out-of-band, when the spurious interfering channels appear at wavelengths spectrally separated from λ, or as in-band crosstalk, when they are equal to λ. For the same amount of crosstalk power, this latter situation is much more critical in terms of overall performance [16]. Both types of crosstalk translate into a power penalty at receivers dependent on the amount of interfering power.

For out-of-band crosstalk, also called incoherent crosstalk, the contribution from adjacent wavelength channels X_A is usually higher than the contribution from non-adjacent channels X_{NA}. Following the formalism presented in [17], the overall crosstalk relative power level, expressed in dimensionless linear units, can be approximated as follows

$$X(w) = 2X_A + (w - 3)X_{NA}, \tag{1.1}$$

where $X(w)$ is the total amount of crosstalk power present on a given port, normalized to the useful signal power out of that port; w is the number of wavelength channels, which is typically equal to the number n of ports of the device. Out-of-band crosstalk is present on any WDM filtering device, such as WDM