Optical Switching: Switch Fabrics, Techniques, and Architectures

Georgios I. Papadimitriou, *Senior Member, IEEE*, Chrisoula Papazoglou, and Andreas S. Pomportsis, *Member, IEEE*

(Invited Tutorial)

***Abstract*—The switching speeds of electronics cannot keep up with the transmission capacity offered by optics. All-optical switch fabrics play a central role in the effort to migrate the switching functions to the optical layer. Optical packet switching provides an almost arbitrary fine granularity but faces significant challenges in the processing and buffering of bits at high speeds. Generalized multiprotocol label switching seeks to eliminate the asynchronous transfer mode and synchronous optical network layers, thus implementing Internet protocol over wavelength-division multiplexing. Optical burst switching attempts to minimize the need for processing and buffering by aggregating flows of data packets into bursts. In this paper, we present an extensive overview of the current technologies and techniques concerning optical switching.**

***Index Terms*—Generalized multiprotocol label switching (GMPLS), optical burst switching (OBS), optical packet switching, optical switch fabrics, optical switching.**

I. Introduction

THE UNPRECEDENTED demand for optical network capacity has fueled the development of long-haul optical network systems which employ wavelength-division multiplexing (WDM) to achieve tremendous capacities. Such systems transport tens to hundreds of wavelengths per fiber, with each wavelength modulated at 10 Gb/s or more [2]. Up to now, the switching burden in such systems has been laid almost entirely on electronics. In every switching node, optical signals are converted to electrical form (O/E conversion), buffered electronically, and subsequently forwarded to their next hop after being converted to optical form again (E/O conversion). Electronic switching is a mature and sophisticated technology that has been studied extensively. However, as the network capacity increases, electronic switching nodes seem unable to keep up. Apart from that, electronic equipment is strongly dependent on the data rate and protocol, and thus, any system upgrade results in the addition and/or replacement of electronic switching equipment. If optical signals could be switched without conversion to electrical form, both of these drawbacks would be eliminated. This is the promise of optical switching.

The main attraction of optical switching is that it enables routing of optical data signals without the need for conversion to electrical signals and, therefore, is independent of data rate and data protocol. The transfer of the switching function from electronics to optics will result in a reduction in the network equipment, an increase in the switching speed, and thus network throughput, and a decrease in the operating power. In addition, the elimination of E/O and O/E conversions will result in a major decrease in the overall system cost, since the equipment associated with these conversions represents the lion's share of cost in today's networks.

Up to now, the limitations of optical component technology, i.e., the lack of processing at bit level and the lack of efficient buffering in the optical domain, have largely limited optical switching to facility management applications. Several solutions are currently under research; the common goal for all researchers is the transition to switching systems in which optical technology plays a more central role.

The three main approaches that seem promising for the gradual migration of the switching functions from electronics to optics are optical packet switching (OPS), generalized multiprotocol label switching (GMPLS), and optical burst switching (OBS). While GMPLS provides bandwidth at a granularity of a wavelength, OPS can offer an almost arbitrary fine granularity, comparable to currently applied electrical packet switching, and OBS lies between them.

This paper is outlined as follows. In Section II, optical switch fabrics, the core of an optical switching system, are presented. Section III presents an overview of optical packet switching, while Sections IV and V present GMPLS and OBS, respectively.

II. Optical Switches

A. Applications of Switches

1) Optical Cross-Connects: A very important application of optical switches is the provisioning of lightpaths. A lightpath is a connection between two network nodes that is set up by assigning a dedicated wavelength to it on its link in its path [1]. In this application, the switches are used inside optical cross-connects (OXCs) to reconfigure them to support new lightpaths. OXCs are the basic elements for routing optical signals in an optical network or system [4]; OXCs groom and optimize transmission data paths [5]. Optical switch requirements for OXCs include

1) scalability;
2) high-port–count switches;
3) the ability to switch with high reliability, low loss, good uniformity of optical signals independent of path length;
4) the ability to switch to a specific optical path without disrupting the other optical paths.

Manuscript received May 27, 2002; revised September 30, 2002.

The authors are with the Department of Informatics, Aristotle University, 54006 Thessaloniki, Greece (e-mail: gp@csd.auth.gr).

Digital Object Identifier 10.1109/JLT.2003.808766

Most of the cross-connects that are currently used in networks use an electrical core for switching where the optical signals are first converted to electrical signals, which are then switched by electrical means and finally converted back to optical signals. This type of switching is referred to as O/E/O switching. This approach features a number of disadvantages. First, the switching speed of electronics cannot keep up with the capacity of optics. Electronic asynchronous transfer mode (ATM) switches and Internet protocol (IP) routers can be used to switch data using the individual channels within a WDM link, but this approach implies that tens or hundreds of switch interfaces must be used to terminate a single link with a large number of channels [49]. Second, O/E/O switching is neither data-rate nor data-format transparent. When the data rate increases, the expensive transceivers and electrical switch core have to be replaced [4].

All-optical cross-connects (OOO cross-connects) switch data without any conversions to electrical form. The core of an OOO cross-connect is an optical switch that is independent of data rate and data protocol, making the cross-connect ready for future data-rate upgrades [4]. Other advantages of OOO cross-connects include reductions in cost, size, and complexity. On the other side, even a scalable and data rate/protocol transparent network is useless if it cannot be managed accordingly. Unfortunately, invaluable network management functions (e.g., performance monitoring and fault isolation) cannot, up to now, be implemented entirely in the optical domain, because of the two major limitations of optical technology (lack of memory and bit processing). Another disadvantage of OOO cross-connects is that they do not allow signal regeneration with retiming and reshaping (3R). This limits the distances that can be traveled by optical signals.

Opaque cross-connects are a compromise between O/E/O and OOO approaches. Opaque cross-connects are mostly optical at the switching fabric but still rely on a limited subset of surrounding electronics to monitor system integrity [5]. Here, the optical signal is converted into electrical signals and then again to optical. The signals are switched in the optical domain and then converted to electrical and finally back to optical again. This option may still improve the performance of the cross-connect, since the optical switch core does not have the bandwidth limitations and power consumption of an electrical switch core. Opaque optical cross-connects allow the options of wavelength conversion, combination with an electrical switch core, quality of service (QoS) monitoring, and signal regeneration, all within the cross-connect switch. However, since there are O/E and E/O conversions, the data-rate and data- format transparency is lost [4].

2) Protection Switching: Protection switching allows the completion of traffic transmission in the event of system or network-level errors. Optical protection switching usually requires optical switches with smaller port counts of 1×2 or 2×2. Protection switching requires switches to be extremely reliable, since sometimes these switches are single points of failure in the network. Protection schemes typically involve several steps that must be taken in order to determine the origin and nature of the failure, to notify other nodes, etc. These processes take longer than the optical switch and thus relax the requirements on the switching speed, which is important but not critical.

3) Optical Add/Drop Multiplexing: Optical add/drop multiplexers (OADMs) residing in network nodes insert (add) or extract (drop) optical channels (wavelengths) to or from the optical transmission stream. Using an OADM, channels in a multiwavelength signal can be added or dropped without any electronic processing. Switches that function as OADMs are wavelength-selective switches, i.e., they can switch the input signals according to their wavelengths.

4) Optical Signal Monitoring: Optical signal monitoring (also referred to as optical spectral monitoring) (OSM) is an important network management operation. OSM receives a small optically tapped portion of the aggregated WDM signal, separates the tapped signal into its individual wavelengths, and monitors each channel's optical spectra for wavelength accuracy, optical power levels, and optical crosstalk.

The size of the optical switch that is used for signal monitoring is chosen based on the system wavelength density and the desired monitoring thoroughness. It is important in the OSM application, because the tapped optical signal is very low in optical signal power, that the optical switch employed has a high extinction ratio (low interference between ports), low insertion loss, and good uniformity [5].

5) Network Provisioning: Network provisioning occurs when new data routes have to be established or existing routes need to be modified. A network switch should carry out reconfiguration requests over time intervals on the order of a few minutes. However, in many core networks today, provisioning for high-capacity data pipes requires a slow manual process, taking several weeks or longer. High-capacity reconfigurable switches that can respond automatically and quickly to service requests can increase network flexibility, and thus bandwidth and profitability.

B. Optical Switch Fabrics

In the effort to extend optics from transmission to switching, all-optical switching fabrics play a central role. These devices allow switching directly in the optical domain, avoiding the need for several O/E/O conversions. Most solutions for all-optical switching are still under study. Given the wide range of possible applications for these devices, it seems reasonable to foresee that there will not be a single winning solution [6]. Before presenting the details of the optical switching technologies available today, we discuss, in brief, the parameters that we take into account when evaluating an optical switch [1].

The most important parameter of a switch is the switching time. Different applications have different switching time requirements. Other important parameters of a switch follow.

1) *Insertion loss*: This is the fraction of signal power that is lost because of the switch. This loss is usually measured in decibels and must be as small as possible. In addition, the insertion loss of a switch should be about the same for all input–output connections (loss uniformity).
2) *Crosstalk*: This is the ratio of the power at a specific output from the desired input to the power from all other inputs.

3) *Extinction ratio* (ON–OFF switches): This is the ratio of the output power in the on-state to the output power in the off-state. This ratio should be as large as possible.
4) *Polarization-dependent loss (PDL)*: If the loss of the switch is not equal for both states of polarization of the optical signal, the switch is said to have polarization-dependent loss. It is desirable that optical switches have low PDL.

Other parameters that are taken into account include reliability, energy usage, scalability, and temperature resistance. The term *scalability* refers to the ability to build switches with large port counts that perform adequately. It is a particularly important concern.

The main optical switching technologies available today follow [5], [6].

1) Optomechanical Switches: Optomechanical technology was the first commercially available for optical switching. In optomechanical switches, the switching function is performed by some mechanical means. These mechanical means include prisms, mirrors, and directional couplers. Mechanical switches exhibit low insertion losses, low polarization-dependent loss, low crosstalk, and low fabrication cost. Their switching speeds are in the order of a few milliseconds, which may not be acceptable for some types of applications. Another disadvantage is the lack of scalability. As with most mechanical components, long-term reliability is also of some concern. Optomechanical switch configurations are limited to 1×2 and 2×2 port sizes. Larger port counts can only be obtained by combining several 1×2 and 2×2 switches, but this increases cost and degrades performance. Optomechanical switches are mainly used in fiber protection and very-low-port-count wavelength add/drop applications.

2) Microelectromechanical System Devices: Although microelectromechanical system (MEMS) devices can be considered as a subcategory of optomechanical switches, they are presented separately, mainly because of the great interest that the telecommunications industry has shown in them, but also because of the differences in performance compared with other optomechanical switches. MEMS use tiny reflective surfaces to redirect the light beams to a desired port by either ricocheting the light off of neighboring reflective surfaces to a port or by steering the light beam directly to a port [5].

One can distinguish between two MEMS approaches for optical switching: two-dimensional (2-D), or digital, and three–dimensional (3–D), or analog, MEMS [4]. In 2-D MEMS, the switches are digital, since the mirror position is bistable (ON or OFF), which makes driving the switch very straightforward. Fig. 1 shows a top view of a 2-D MEMS device with the microscopic mirrors arranged in a crossbar configuration to obtain cross-connect functionality. Collimated light beams propagate parallel to the substrate plane. When a mirror is activated, it moves into the path of the beam and directs the light to one of the outputs, since it makes a 45° angle with the beam. This arrangement also allows light to be passed through the matrix without hitting a mirror. This additional functionality can be used for adding or dropping optical channels (wavelengths). The tradeoff for the simplicity of the mirror control in a 2-D MEMS switch is optical loss. While the path length grows linearly with the number of ports, the optical loss grows rapidly. Commercially available products feature a maximum insertion loss of 3.7 dB for an 8×8 switch, 5.5 dB for 16×16, and 7.0 for 32×32 [53]. Therefore, 2-D architectures are found to be impractical beyond 32-input and 32-output ports. While multiple stages of 32×32 switches can theoretically form a 1000-port switch, high optical losses also make such an implementation impractical [2]. High optical losses could be compensated by optical amplification, but this will increase the overall system cost. Apart from cost considerations, optical amplifiers are by no means ideal devices. First, optical amplifiers introduce noise, in addition to providing gain. Second, the gain of the amplifier depends on the total input power. For high-input powers, the amplifier tends to saturate and the gain drops. This can cause undesirable power transients in networks. Finally, although optical amplifiers are capable of amplifying many wavelength channels simultaneously, they do not amplify all channels equally, i.e., their gain is not flat over the entire passband [1].

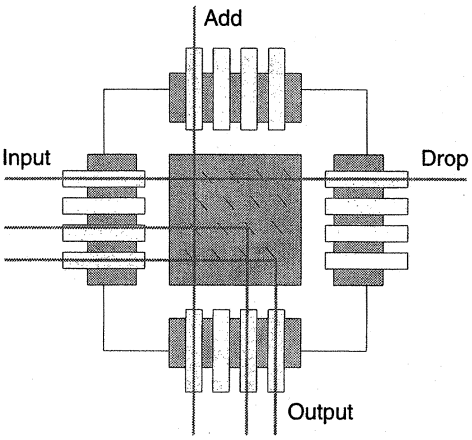

Fig. 1. 2-D MEMS technology.

In 3-D MEMS, there is a dedicated movable mirror for each input and each output port. A connection path is established by tilting two mirrors independently to direct the light from an input port to a selected output port. Mirrors operate in an analog mode, tilting freely about two axes [1]. This is a most promising technology for very-large-port-count OXC switches with >1000 input and output ports. A drawback of this approach is that a complex (and very expensive) feedback system is required to maintain the position of mirrors (to stabilize the insertion loss) during external disturbances or drift.

The actuation forces that move the parts of the switch may be electrostatic, electromagnetic, or thermal. Magnetic actuation offers the benefit of large bidirectional (attractive and repulsive) linear force output but requires a complex fabrication process and electromagnetic shielding. Electrostatic actuation is the preferred method, mainly because of the relative ease of fabrication and integration and because it allows extremely low-power dissipation.

MEMS technology enables the fabrication of actuated mechanical structures with fine precision that are barely visible to the human eye. MEMS devices are, by nature, compact and consume low power. A batch fabrication process allows high-volume production of low-cost devices, where hundreds or thou-

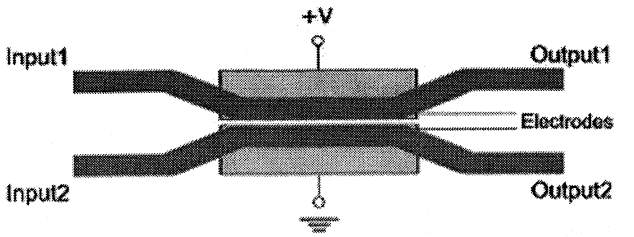

Fig. 2. An electrooptic directional coupler switch.

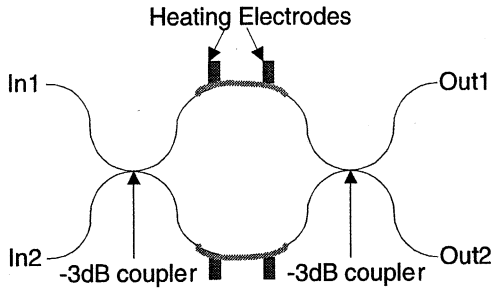

Fig. 3. Scheme of a 2×2 interferometric switch.

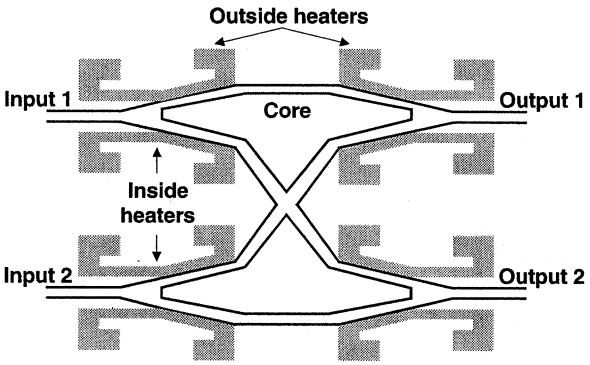

Fig. 4. Scheme of a 2×2 digital optical switch.

sands of devices can be built on a single silicon wafer. Optical MEMS is a promising technology to meet the optical switching need for large-port-count high-capacity OXCs. Potential benefits of an all-optical MEMS-based OXC include scalability, low loss, short switching time, low power consumption, low crosstalk and polarization effects, and independence of wavelength and bit rate [1]. Other applications for MEMS include wavelength add/drop multiplexing, optical service monitoring, and optical protection switching. Challenges concerning MEMS include mirror fabrication, optomechanical packaging, mirror control algorithm, and implementation.

3) Electrooptic Switches: A 2×2 electrooptic switch [1], [5] uses a directional coupler whose coupling ratio is changed by varying the refractive index of the material in the coupling region. One commonly used material is lithium niobate $(LiNbO_3)$. A switch constructed on a lithium niobate waveguide is shown in Fig. 2. An electrical voltage applied to the electrodes changes the substrate's index of refraction. The change in the index of refraction manipulates the light through the appropriate waveguide path to the desired port.

An electrooptic switch is capable of changing its state extremely rapidly, typically in less than a nanosecond. This switching time limit is determined by the capacitance of the electrode configuration. Electrooptic switches are also reliable, but they pay the price of high insertion loss and possible polarization dependence. Polarization independence is possible but at the cost of a higher driving voltage, which in turn limits the switching speed. Larger switches can be realized by integrating several 2×2 switches on a single substrate. However, they tend to have a relatively high loss and PDL and are more expensive than mechanical switches.

4) Thermooptic Switches: The operation of these devices [6] is based on the thermooptic effect. It consists in the variation of the refractive index of a dielectric material, due to temperature variation of the material itself. There are two categories of thermooptic switches: interferometric and digital optical switches.

Interferometric switches are usually based on Mach–Zehnder interferometers. These devices, as shown in Fig. 3, consist of a 3-dB coupler that splits the signal into two beams, which then travel through two distinct arms of same length, and of a second 3-dB coupler, which merges and finally splits the signal again. Heating one arm of the interferometer causes its refractive index to change. Consequently, a variation of the optical path of that arm is experienced. It is thus possible to vary the phase difference between the light beams, by heating one arm of the interferometer. Hence, as interference is constructive or destructive, the power on alternate outputs is minimized or maximized. The output port is thus selected.

Digital optical switches [6] are integrated optical devices generally made of silica on silicon. The switch is composed of two interacting waveguide arms through which light propagates. The phase error between the beams at the two arms determines the output port. Heating one of the arms changes its refractive index, and the light is transmitted down one path rather than the other. An electrode through control electronics provides the heating. A 2×2 digital optical switch is shown in Fig. 4 ([7]).

Thermooptical switches are generally small in size but have the drawback of having high-driving-power characteristics and issues in optical performance. The disadvantages of this technology include limited integration density (large die area) and high-power dissipation. Most commercially available switches of this type require forced air cooling for reliable operation. Optical performance parameters, such as crosstalk and insertion loss, may be unacceptable for some applications. On the positive side, this technology allows the integration of variable optical attenuators and wavelength selective elements (arrayed waveguide gratings) on the same chip with the same technology [4].

5) Liquid-Crystal Switches: The liquid-crystal state is a phase that is exhibited by a large number of organic materials over certain temperature ranges. In the liquid-crystal phase, molecules can take up a certain mean relative orientation, due to their permanent electrical dipole moment. It is thus possible, by applying a suitable voltage across a cell filled with liquid-crystal material, to act on the orientation of the molecules. Hence, optical properties of the material can be altered. Liquid-crystal optical switches [4], [6] are based on the change of polarization state of incident light by a liquid

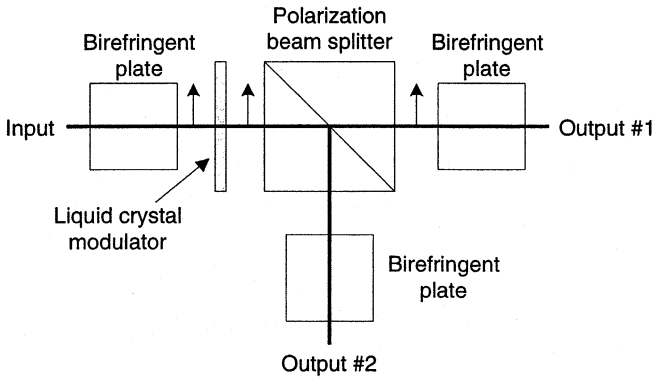

Fig. 5. Scheme of 1×2 liquid-crystal optical switch.

crystal as a result of the application of an electric field over the liquid crystal. The change of polarization in combination with polarization selective beam splitters allows optical space switching. In order to make the devices polarization insensitive, some kind of polarization diversity must be implemented that makes the technology more complex. Polarization diversity schemes attempt to make devices polarization insensitive by treating each polarization mode differently. The input signal is decomposed into its TE and TM components. Each component is treated separately in the switch. At the output, the TE and TM components are recombined.

A 1×2 liquid-crystal optical switch structure is shown in Fig. 5. The principle of operation is as follows [8]: the birefringent plate at the input port manipulates the polarization states to the desired ones. Birefringent materials have different refractive indexes along two different directions (for example, the x and y axes). Without applying a bias, the input signal passes through the liquid-crystal cell and polarization beam splitter with the same polarization. By applying a voltage on the liquid-crystal spatial modulator, molecules rotate the polarizations of the signal passing through them. With sufficient voltage, the signal polarizations rotate to the orthogonal ones and the polarization beam splitter reflects the signal to the other output port.

Liquid-crystal switches have no moving parts. They are very reliable, and their optical performance is satisfactory, but they can be affected by extreme temperatures if not properly designed.

6) Bubble Switches : Bubble switches [6], [11] can be classified as a subset of thermooptical technology, since their operation is also based on heating and cooling of the substrate. However, the behavior of bubble switches when heated is different from other thermooptic switches, which were described previously.

This technology is based on the same principle as for ink-jet printers. The switch is made up of two layers: a silica bottom layer, through which optical signals travel, and a silicon top layer, containing the ink-jet technology. In the bottom layer, two series of waveguides intersect each other at an angle of around 120°. At each cross-point between two guides, a tiny hollow is filled in with a liquid that exhibits the same refractive index of silica, in order to allow propagation of signals in normal conditions. When a portion of the switch is heated, a refractive index change is caused at the waveguide junctions. This effect results in the generation of tiny bubbles. Thus, a light beam travels straight through the guide, unless the guide is interrupted by a bubble placed in one of the hollows at the cross-points. In this case, light is deflected into a new guide, crossing the path of the previous one.

This technology relies on proven ink-jet printer technology and can achieve good modular scalability. However, for telecom environments, uncertainty exists about long-term reliability, thermal management, and optical insertion losses.

7) Acoustooptic Switches: The operation of acoustooptic switches [9], [10] is based on the acoustooptic effect, i.e., the interaction between sound and light. The principle of operation of a polarization-insensitive acoustooptic switch is as follows [10]. First, the input signal is split into its two polarized components (TE and TM) by a polarization beam splitter (Fig. 6). Then, these two components are directed to two distinct parallel waveguides. A surface acoustic wave is subsequently created. This wave travels in the same direction as the lightwaves. Through an acoustooptic effect in the material, this forms the equivalent of a moving grating, which can be phase-matched to an optical wave at a selected wavelength. A signal that is phase-matched is "flipped" from the TM to the TE mode (and vice versa), so that the polarization beam splitter that resides at the output directs it to the lower output. A signal that was not phase-matched exits on the upper output.

If the incoming signal is multiwavelength, it is even possible to switch several different wavelengths simultaneously, as it is possible to have several acoustic waves in the material with different frequencies at the same time. The switching speed of acoustooptic switches is limited by the speed of sound and is in the order of microseconds.

8) Semiconductor Optical Amplifier Switches: Semiconductor optical amplifiers (SOAs) [1], [12] are versatile devices that are used for many purposes in optical networks. An SOA can be used as an ON–OFF switch by varying the bias voltage. If the bias voltage is reduced, no population inversion is achieved, and the device absorbs input signals. If the bias voltage is present, it amplifies the input signals. The combination of amplification in the on-state and absorption in the off-state makes this device capable of achieving very high extinction ratios. Larger switches can be fabricated by integrating SOAs with passive couplers. However, this is an expensive component, and it is difficult to make it polarization independent [1].

Table I compares optical switching technologies. All figures were derived from data sheets for commercially available products.

C. Large Switches

Switch sizes larger than 2×2 can be realized by appropriately cascading small switches. The main considerations in building large switches are the following [1].

1) Number of Small Switches Required: Optical switches are made by cascading 2×2 or 1×2 switches, and thus, the cost is, to some extent, proportional to the number of such switches needed. However, this is only one of the factors that affect the cost. Other factors include packaging, splicing, and ease of fabrication.

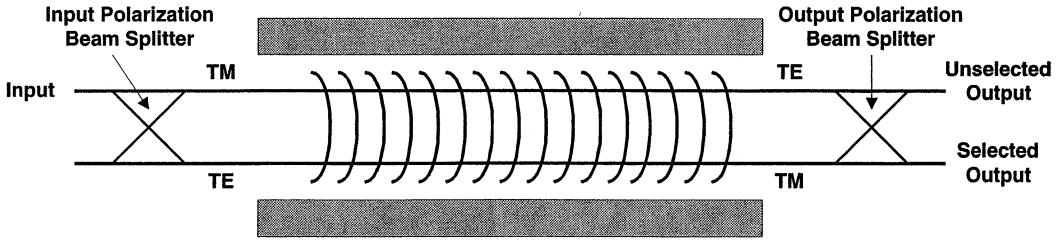

Fig. 6. Schematic of a polarization independent acoustooptic switch.

TABLE I
COMPARISON OF OPTICAL SWITCHING TECHNOLOGIES

	Insertion Loss (dB)	Crosstalk (dB)	PDL (dB)	Switching Time
Optomechanical 8×8 [54]	0.5	-55	0.07	4ms
MEMS 8×8 [53]	0.2-3.7	-50	0.4	12ms
Electro-optic 8×8 [55]	9	-30		5ns
Thermo-optic 8×8 [56]	8		0.5	3ms
Liquid crystals 2×2 [57]	1.4	-50	0.2	5ms
Bubble 2×2 [58]	2.5-7.5	-50	0.3	10ms
Acousto-optic 1×N [59]	6	-35		3μs

2) Loss Uniformity: Switches may have different losses for different combinations of input and output ports. This situation is exacerbated for large switches. A measure of the loss uniformity can be obtained by considering the minimum and maximum number of switch elements in the optical path for different input and output combinations (this number should be nearly constant).

3) Number of Crossovers: Large optical switches are sometimes fabricated by integrating multiple switch elements on a single substrate. Unlike integrated electronic circuits (ICs), where connections between the various components can be made at multiple layers, in integrated optics, all these connections must be made in a single layer by means of waveguides. If the paths of two waveguides cross, two undesirable effects are introduced: power loss and crosstalk. In order to have acceptable loss and crosstalk performance for the switch, it is thus desirable to minimize, or completely eliminate, such waveguide crossovers.

4) Blocking Characteristics: In terms of the switching function achievable, switches are of two types: blocking or nonblocking in the presence of other lightpaths. A switch is said to be nonblocking if an unused input port can be connected to any unused output port. Thus, a nonblocking switch is capable of realizing every interconnection pattern between the inputs and the outputs. If some interconnection pattern cannot be realized, the switch is said to be blocking. Most applications require nonblocking switches. However, even nonblocking switches can be further distinguished in terms of the effort needed to achieve the nonblocking property. A switch is said to be wide-sense nonblocking if any unused input can be connected to any unused output, without requiring any existing connection to be rerouted. In addition, a switch that is nonblocking, regardless of the connection rule that is used, is said to be strict-sense nonblocking.

A nonblocking switch that may require rerouting of connections to achieve the nonblocking property is said to be rearrangeably nonblocking. Rerouting of connections may or may not be acceptable, depending on the application, since existing connections must be interrupted, at least briefly, in order to switch it to a different path. The advantage of rearrangeably nonblocking switch architectures is that they use fewer small switches to build a larger switch of a given size, compared with the wide-sense nonblocking switch architectures.

While rearrangeably nonblocking architectures use fewer switches, they require a more complex control algorithm to set up connections. Since optical switches are not very large, the increased complexity may be acceptable. The main drawback of rearrangeably nonblocking switches is that many applications will not allow existing connections to be disrupted, even temporarily, to accommodate a new connection.

Usually, there is a tradeoff between these different aspects. The most popular architectures for building large switches [1] are the crossbar.

1) Crossbar: An $n \times n$ crossbar is made of n^2 2×2 switches. The interconnection between the inputs and the outputs is achieved by appropriately setting the states of the 2×2 switches. The connection rule that is used states that to

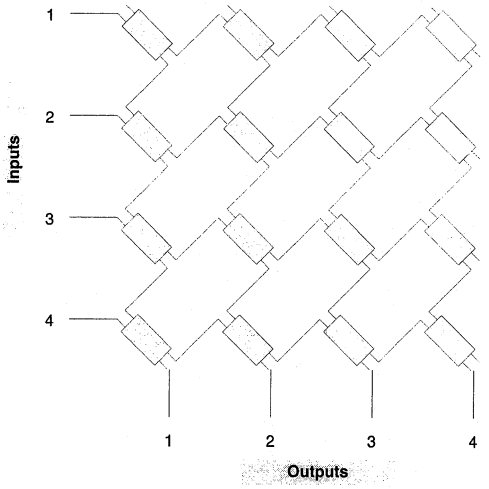

Fig. 7. 4×4 switch realized using 16 2×2 switches.

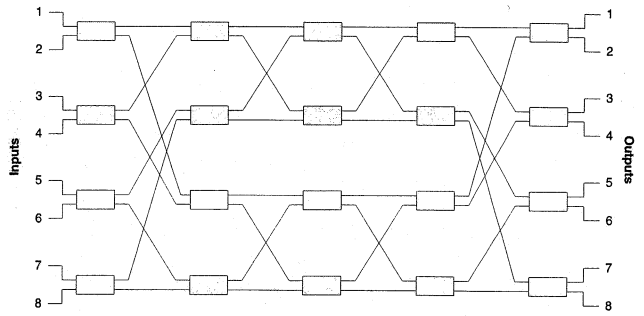

Fig. 8. Rearrangeably nonblocking 8×8 switch realized using 20 2×2 switches interconnected in the Beneš architecture.

connect input i to output j, the path taken traverses the 2×2 switches in row i until it reaches column j and then traverses the switches in column j until it reaches output j.

The crossbar architecture (Fig. 7) is wide-sense nonblocking; therefore, as long as the connection rule mentioned previously is used. The shortest path length is 1, and the longest path length is $2n-1$, and this is one of the main drawbacks of the crossbar architecture. The switch can be fabricated without any crossovers.

2) Beneš : The Beneš architecture [13] (Fig. 8) is a rearrangeably nonblocking switch architecture and is one of the most efficient switch architectures in terms of the number of 2×2 switches it uses to build larger switches. An $n \times n$ Beneš switch requires $(n/2)(2\log_2 n - 1)$ 2×2 switches, n being a power of 2. The loss is the same through every path in the switch—each path goes through $2\log_2 n - 1$ 2×2 switches. Its two main drawbacks are that it is not wide-sense nonblocking and that a number of waveguide crossovers are required, making it difficult to fabricate in integrated optics.

3) Spanke–Beneš (n-Stage Planar Architecture): This switch architecture (Fig. 9) is a good compromise between the crossbar and Beneš switch architectures. It is rearrangeably nonblocking and requires $n(n-1)/2$ switches. The shortest path length is $n/2$, and the longest path length is n. There are no crossovers. Its main drawbacks are that it is not wide-sense nonblocking and that the loss is not uniform.

4) Spanke: This architecture (Fig. 10) is suitable for building large nonintegrated switches. An $n \times n$ switch is made by combining n $1 \times n$ switches, along with n $n \times 1$ switches. The architecture is strict-sense nonblocking and requires $2n(n-1)$ 1×2 switches, and each path has length $2\log_2 n$.

III. Optical Packet Switching

Using pure WDM only provides granularity at the level of one wavelength. If data at a capacity of a fraction of a wavelength's granularity is to be carried, capacity will be wasted. With optical packet switching, packet streams can be multiplexed together statistically, making more efficient use of capacity and providing increased flexibility over pure WDM [14]. The wavelength dimension is also used inside the optical packet switch in order to allow the optical buffers to be used efficiently and the switch throughput to be increased.

Packet switches analyze the information contained in the packet headers and thus determine where to forward the packets. Optical packet-switching technologies enable the fast allocation of WDM channels in an on-demand fashion with fine granularities (microsecond time scales). An optical packet switch can cheaply support incremental increases of the transmission bit rate so that frequent upgrades of the transmission layer capacity can be envisaged to match increasing bandwidth demand with a minor impact on switching nodes [15]. In addition, optical packet switching offers high-speed, data rate/format transparency, and configurability, which are some of the important characteristics needed in future networks supporting different forms of data [16].

A. Issues Concerning Optical Packet Switching

Optical packet-switched networks can be divided into two categories: slotted (synchronous) and unslotted (asynchronous). In a slotted network, all packets have the same size. They are placed together with the header inside a fixed time slot, which has a longer duration than the packet and header to provide guard time. In a synchronous network, packets arriving at the input ports must be aligned in phase with a local clock reference [16]. Maintaining synchronization is not a simple task in the optical domain. Assuming an Internet environment, fixed-length packets imply the need to segment IP datagrams at the edge of the network and reassemble them at the other edge. This can be a problem at very high speeds. For this reason, it is worth considering asynchronous operation with variable-length packets [14], [17].

Packets in an unslotted network do not necessarily have the same size. Packets in an asynchronous network arrive and enter the switch without being aligned. Therefore, the switch action could take place at any point in time. The behavior of packets in an unslotted network is not predictable. This leads to an increased possibility of packet contention and therefore impacts negatively on the network throughput. Asynchronous operation also leads to an increased packet loss ratio. However, the use of the wavelength domain for contention resolution, as is described subsequently, can counteract this. On the other hand, unslotted networks feature a number of advantages, such as increased robustness and flexibility, as well as lower cost and ease of setup. Thereby, a good traffic performance is attained, while the use of complicated packet alignment units is avoided [18].

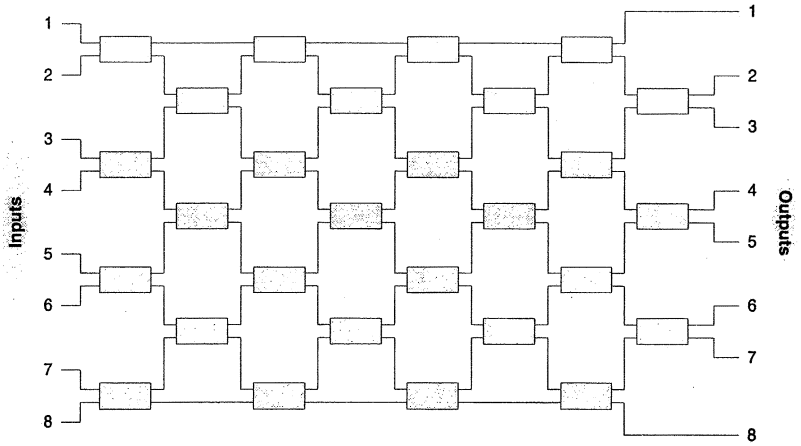

Fig. 9. Rearrangeably nonblocking 8×8 switch realized using 28 2×2 switches and no waveguide crossovers interconnected in the n-stage planar architecture.

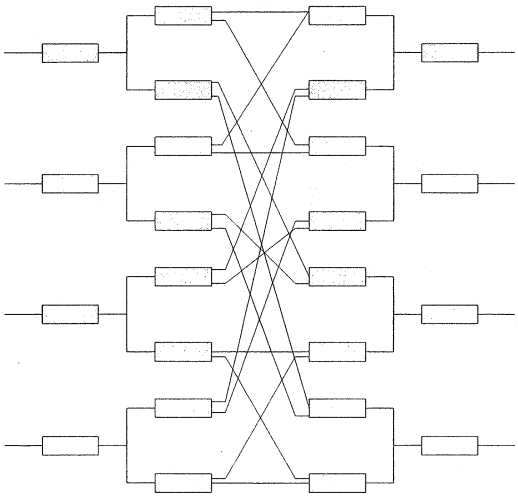

Fig. 10. Strict-sense nonblocking 4×4 switch realized using 24 $1 \times 2/2 \times 1$ switches interconnected in the Spanke architecture.

Packets traveling in a packet-switched network experience variant delays. Packets traveling on a fiber can experience different delays, depending on factors such as fiber length, temperature variation, and chromatic dispersion [16]. *Chromatic dispersion* is the term given to the phenomenon by which different spectral components of a pulse travel at different velocities [1]. In other words, because of chromatic dispersion, packets that are transmitted on different wavelengths experience different propagation delays. The use of dispersion-compensating fiber alleviates the effects of chromatic dispersion. The packet propagation speed is also affected by temperature variations. The sources of delay variations described so far can be compensated statically and not dynamically (on a packet-by-packet basis) [16].

The delay variations mentioned previously were delays that packets experience while they are transmitted between network nodes. The delays that packets experience in switching nodes are also not fixed. The contention resolution scheme, and the switch fabric, greatly affect the packet delay. In a slotted network that uses fiber delay lines (FDLs) as optical buffers, a packet can take different paths with unequal lengths within the switch fabric [16].

Packets that arrive in a packet-switching node are directed to the switch's input interface. The input interface aligns the packet so that they will be switched correctly (assuming the network operates in a synchronous manner) and extracts the routing information from the headers. This information is used to control the switching matrix. The switching matrix performs the switching and buffering functions. The control is electronic, since optical logic is in too primitive a state to permit optical control currently. After the switching, packets are directed to the output interface, where their headers are rewritten. The operating speed of the control electronics places an upper limit on the switch throughput. For this reason, it is imperative that the packet switch control scheme and the packet scheduling algorithms are kept as simple as possible.

The header and payload can be transmitted serially on the same wavelength. Guard times must account for payload position jitter and are necessary before and after the payload to prevent damages during header erasure or insertion [19]. Although there are various techniques to detect and recognize packet headers at gigabit-per-second speed, either electronically or optically [20], [21], it is still difficult to implement electronic header processors operating at such high speed as to switch packets on the fly at every node [16].

Several solutions have been proposed for this problem. One of these suggestions employs subcarrier multiplexing. In this approach, the header and payload are multiplexed on the same wavelength, but the payload data is encoded at the baseband, while header bits are encoded on a properly chosen subcarrier frequency at a lower bit rate. This enables header retrieval without the use of an optical filter. The header can be retrieved using a conventional photodetector. This approach features several advantages, such as the fact that the header interpretation process can take up the whole payload transmission time, but also puts a possible limit on the payload data rate. If the payload data rate is increased, the baseband will expand and might eventually overlap with the subcarrier frequency, which is limited by the microwave electronics.

According to another approach, the header and the payload are transmitted on separate wavelengths. When the header needs to be updated, it is demultiplexed from the payload and processed electronically. This approach suffers from fiber

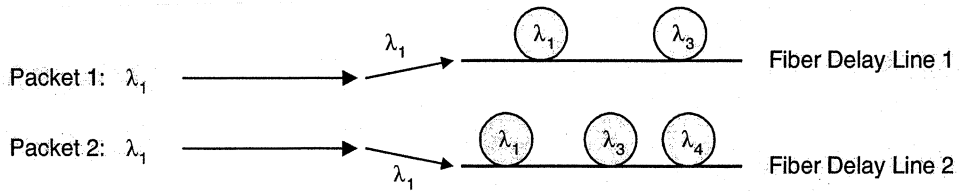

Fig. 11. Assignment of packets to FDLs without the use of tunable optical wavelength converters.

dispersion, which separates the header and payload as the packet propagates through the network. Subcarrier multiplexed headers have far less dispersion problems, since they are very close to the baseband frequency.

B. Contention Resolution

Two major difficulties prevail in optical packet switching: there is currently no capability of bit-level processing in the optical domain, and there is no efficient way to store information in the optical domain indefinitely. The former issue concerns the process of reading and interpreting the packet headers, while the latter concerns the way packet contentions are resolved in an optical network. Contentions occur in the network switches when two or more packets have to exploit the same resource, for example, when two packets must be forwarded to the same output channel at the same time. The adopted solutions to solve these contentions are a key aspect in packet-switched networks, and they can heavily affect the overall network performance. The optical domain offers new ways to solve contentions but does not allow the implementation of methods that are widely used in networks today. Three methods for contention resolution are analyzed in the following: buffering, deflection routing, and wavelength conversion.

1) Buffering: The simplest solution to overcome the contention problem is to buffer contending packets, thus exploiting the time domain. This technique is widely used in traditional electronic packet switches, where packets are stored in the switch's random access memory (RAM) until the switch is ready to forward them. Electronic RAM is cheap and fast. On the contrary, optical RAM does not exist. FDLs are the only way to "buffer" a packet in the optical domain. Contending packets are sent to travel over an additional fiber length and are thus delayed for a specific amount of time.

Optical buffers are either single-stage or multistage, where the term stage represents one or more parallel continuous piece of delay line [16]. Optical buffer architectures can be further categorized into feed-forward architectures and feedback architectures [1]. In a feedback architecture, the delay lines connect the outputs of the switch to its inputs. When two packets contend for the same output, one of them can be stored in a delay line. When the stored packet emerges at the output of the FDL, it has another opportunity to be routed to the appropriate output. If contention occurs again, the packet is stored again and the whole process is repeated. Although it would appear so, a packet cannot be stored indefinitely in a feedback architecture, because of unacceptable loss. In a feedback architecture, arriving packets can preempt packets that are already in the switch. This allows the implementation of multiple QoS classes.

In the feed-forward architecture, a packet has a fixed number of opportunities to reach its desired output [1]. Almost all the loss that a signal experiences in a switching node is related with the passing through the switch. The feed-forward architecture attenuates all signals almost equally because every packet passes through the same number of switches.

The implementation of optical buffers using FDLs features several disadvantages. FDLs are bulky and expensive. A packet cannot be stored indefinitely on an FDL. Generally, once a packet has entered an FDL, it cannot be retrieved before it emerges on the other side, after a certain amount of time. In other words, FDLs do not have random access capability. Apart from that, optical signals that are buffered using FDLs experience additional quality degradation, since they are sent to travel over extra pieces of fiber. The number of FDLs, as well as their lengths, are critical design parameters for an optical switching system. The number of FDLs required to achieve a certain packet loss rate increases with the traffic load. The length(s) of the FDLs are dictated by the packet duration(s). For the reasons mentioned previously, it is desirable that the need for buffering is minimized. If the network operates in a synchronous manner, the need for buffering is greatly reduced.

The wavelength dimension can be used in combination with optical buffering. The use of the wavelength dimension minimizes the number of FDLs. Assuming that n wavelengths can be multiplexed on a single FDL, each FDL has a capacity of n packets. The more wavelengths, the more packets can be stored on each delay line. Tunable optical wavelength converters (TOWCs) can be used to assign packets to unused wavelengths in the FDL buffers [22].

When TOWCs are not employed, and two packets that have the same wavelength need to be buffered simultaneously, two FDLs are needed to store the packets (Fig. 11). By using TOWCs to assign packets to unused wavelengths in the FDL buffers, a reduction in the number of FDLs in the WDM packet switch is obtained.

As shown in Fig. 12, one of the two packets that have the same wavelength and need to be buffered simultaneously can be converted to another wavelength. Then, both packets can be stored in the same FDL. Packets are assigned to a wavelength on a particular FDL by the buffer control algorithm. The choice of the buffer control algorithm is also a critical decision, since it can greatly affect the packet loss rate. In [23], four different control algorithms are presented and evaluated. It is assumed that packets have variable lengths, which are multiples of a basic time unit (slot) and that new packets arrive only at the beginning of each time slot. The algorithms presented follow.

- *Pure round robin:* Packets are assigned to wavelengths in a round robin fashion. There is no effort to determine whether a

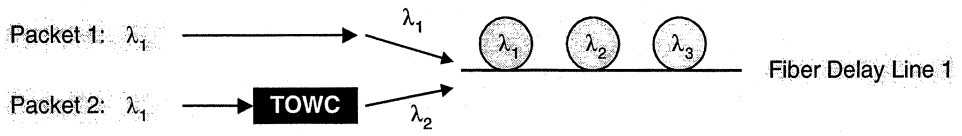

Fig. 12. Assignment of packets to FDLs using tunable optical wavelength converters.

wavelength is available or not. Packets are served in the order of arrival. This algorithm minimizes the control complexity but has a poor performance in terms of packet loss rate and does not utilize the FDL buffer efficiently.

- *Round robin with memory*: This algorithm assigns packets to wavelengths in a round robin fashion but also tracks the occupancy of each wavelength. If an arriving packet is assigned to a wavelength with full occupancy, the algorithm then reassigns the packet to the next available wavelength.
- *Round robin with memory, finding minimum occupancy of the buffer*: This algorithm assigns packets to the least occupied wavelength.
- *Shortest packet first and assign to minimum occupancy buffer*: This algorithm sorts arriving packets according to their length and assigns the shortest packet to the least occupied wavelength.

2) Deflection Routing: This technique resolves contentions by exploiting the space domain. If two or more packets need to use the same output link to achieve minimum distance routing, then only one will be routed along the desired link, while others will be forwarded on paths that may lead to greater than minimum distance routing [16]. The deflected packets may follow long paths to their destinations, thus suffering high delays. In addition, the sequence of packets may be disturbed.

Deflection routing can be combined with buffering in order to keep the packet loss rate below a certain threshold. Deflection routing without the use of optical buffers is often referred to as hot-potato routing. When no buffers are employed, the packet's queuing delay is absent, but the propagation delay is larger than in the buffer solution because of the longer routes that packets take to reach their destination. Simple deflection methods without buffers usually introduce severe performance penalties in throughput, latency, and latency distribution [24].

The most important advantage of the deflection routing method is that it does not require huge efforts to be implemented, neither in terms of hardware components, nor in terms of control algorithms. The effectiveness of this technique critically depends on the network topology; meshed topologies with a high number of interconnections greatly benefit from deflection routing, whereas minor advantages arise from more simple topologies [15]. Moreover, clever deflection rules can lead to an increase in the network throughput. These rules determine which packets will be deflected and where they will be deflected. For example, the alternate link(s) could be found using the second shortest path algorithm. If link utilizations are also taken into account, the packet may be deflected to an underutilized link in order to balance the network load.

When deflection is implemented, a potential problem that may arise is the introduction of routing loops. If no action is taken to prevent loops, then a packet may return to nodes that it has already visited and may remain in the network for an indefinite amount of time. The looping of packets contributes to increased delays and degraded signal quality for the looping packets, as well as an increased load for the entire network [26]. Loops can be avoided by maintaining a hop counter for each deflected packet. When this counter reaches a certain threshold, the packet is discarded. Another approach focuses on the development of deflection algorithms that specifically avoid looping.

3) Wavelength Conversion: The additional dimension that is unique in the field of optics, the wavelength, can be utilized for contention resolution. If two packets that have the same wavelength are addressing the same switch outlet, one of them can be converted to another wavelength using a tunable optical wavelength converter. Only if the wavelengths run out is it necessary to resort to optical buffering. This technique reduces the inefficiency in using the FDLs, particularly in asynchronous switching architectures [17].

By splitting the traffic load on several wavelength channels and by using tunable optical wavelength converters, the need for optical buffering is minimized or even completely eliminated. The authors of [27] consider a WDM packet switch without optical buffers. The network considered operates in a synchronous manner, and the traffic is assumed to be random with a load of 0.8. Results obtained by simulations and calculations show that, when more than 11 WDM channels are used, the packet loss probability is less than 10^{-10} even without any optical buffers [27]. This scheme solves the problem regarding optical buffering. It does, however, require an increase in the number of TOWCs, since for each wavelength channel at a switch input, one tunable wavelength converter is needed. For a 16×16 switch with 11 wavelengths per input, this results in 176 wavelength converters. The size of the space switch, however, is greatly reduced when there are no optical buffers. If wavelength conversion is used in combination with FDLs, the number of required converters is considerably reduced. By using only two FDLs, the number of wavelength channels is reduced from 11 to 4, which results to a considerable decrease of $\sim 64\%$ in the number of converters [27].

In order to reduce the number of converters needed while keeping the packet loss rate low, wavelength conversion must be optimized. Not all packets need to be shifted in wavelength. Decisions must be made concerning the packets that need conversion and the wavelengths to which they will be converted.

C. Packet Switch Architectures

1) General: A general optical WDM packet switch that is employed in a synchronous network consists of three main blocks (Fig. 13) [28].

- *Cell encoder*: Packets arriving at the switch inputs are selected by a demultiplexer, which is followed by a set of tunable optical wavelength converters that address free

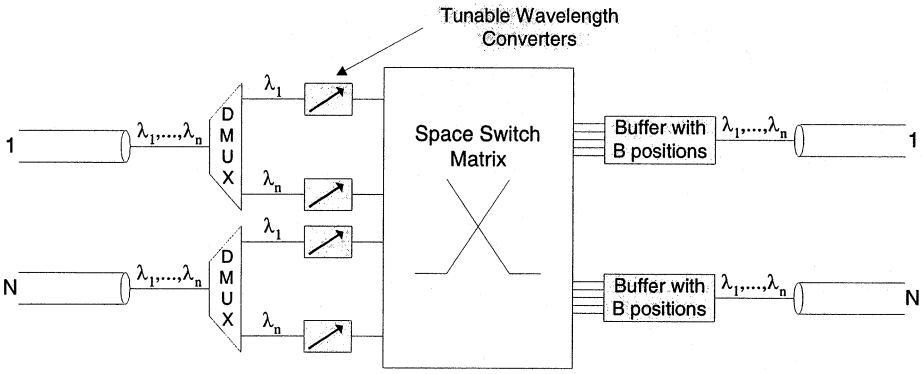

Fig. 13. WDM packet switch.

space in the FDL output buffers. O/E interfaces situated after the demultiplexers extract the header of each packet where the packet's destination is written and thus determine the proper switch outlet. This information is used to control the switch. In addition, optical packet synchronizers must be placed at the switch inlets to assure synchronous operation.

- *Nonblocking space switch*: This switch is used to access the desired outlet as well as appropriate delay line in the output buffer. The size of the space switch in terms of gates is $Nn \times N(B/n + 1)$, where n is the number of wavelengths per fiber, N is the number of input and output fibers, and B is the number of packet positions in the buffer (hence, B/n is the number of FDLs).
- *Buffers*: The switch buffers are realized using FDLs.

The effect of an increase in the number of wavelength channels on the switch throughput is studied in [22]. Simulations that were carried out by the authors of [22] show that the product of the number of fibers in the buffer and the number of wavelength channels $(n \times (B/n + 1))$ remains almost constant when the number of wavelengths is increased. Since the size of the optical switch depends on this product, it is evident that the size of the space switch remains almost constant when the number of wavelength channels is increased. Therefore, by increasing the number of wavelength channels, the throughput of the switch (which is equal to $N \times n \times \rho$ times the channel bit rate, where ρ is the channel load) is increased without increasing the number of gates in the space switch.

The component count for the total switch does not remain constant when more wavelength channels are added. Each channel that is added requires an extra TOWC. Additional simulations show that when tunable optical wavelength converters are employed, the higher allowed burstiness is increased. For example, for a fixed throughput per fiber equal to 0.8 and four wavelength channels, the uses of TOWCs increases the tolerated burstiness from 1.1 to 3.2 [22].

2) Shared Wavelength Converters: In the packet switch architecture described previously, tunable optical wavelength converters were used to handle packet contentions and efficiently access the packet buffers. The use of TOWCs greatly improves the switch performance but results in an increase in the component count and thus cost. In the scheme discussed previously, a wavelength converter is required for each wavelength channel, i.e., for N inputs, each carrying n wavelengths, nN wavelength converters will have to be employed. However, as is noted in [25], only a few of the available TOWCs are simultaneously utilized; this is due to two main reasons.

- Unless a channel load of 100% is assumed, not all channels contain packets at a given instant.
- Not all of the packets contending for the same output line have to be shifted in wavelength because they are already carried by different wavelengths.

These observations suggest an architecture in which the TOWCs are shared among the input channels and their number is minimized so that only those TOWCs strictly needed to achieve given performance requirements are employed.

A bufferless packet switch with shared tunable wavelength converters is shown in Fig. 14. This switch is equipped with a number r of TOWCs, which are shared among the input channels. At each input line, a small portion of the optical power is tapped to the electronic controller, which is not shown in the figure. The switch control unit detects and reads packet headers and drives the space switch matrix and the TOWCs. Incoming packets on each input are wavelength demultiplexed. An electronic control logic processes the routing information contained in each packet header, handles packet contentions, and decides which packets have to be wavelength shifted. Packets not requiring wavelength conversion are directly routed toward the output lines; on the contrary, packets requiring wavelength conversions will be directed to the pool of r TOWCs and, after a proper wavelength conversion, they will reach the output line.

The issue of estimating the number of TOWCs needed to satisfy predefined constraints on the packet loss is addressed in [25]. Packet arrivals are synchronized on a time slot basis and, hence, the number of converters needed at a given time slot depends only on the number of packets arriving at such a slot. The performance of the switch, expressed in terms of packet loss probability depends only on the traffic intensity. Thus, both the converters' dimensioning procedures and the switch performance hold for any type of input traffic statistic. The dimensioning of the converters does not depend on the considered traffic type but only on its intensity.

Converter sharing allows a remarkable reduction of the number of TOWCs with respect to that needed by other switch architectures in which there are as many TOWCs as the input channels. The drawbacks involved in the sharing of TOWCs

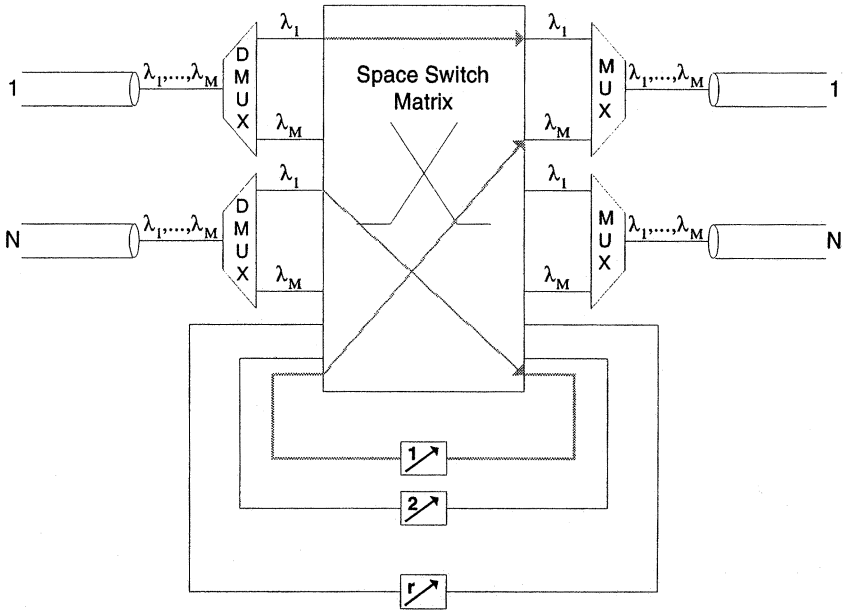

Fig. 14. Packet switch employing shared tunable wavelength converters.

that remain to be dealt with are 1) the enlargement of the switching matrix in order to take into account the sharing of TOWCs and 2) the introduction of an additional attenuation of the optical signal caused by crossing the switching matrix twice.

3) Limited-Range Wavelength Converters: The wavelength converters that were employed in the switch architectures discussed previously were assumed to be capable of converting to and from wavelengths over the full range of wavelengths. In practical systems, a wavelength converter normally has a limited range of wavelength conversion capability. Moreover, a wide range wavelength conversion may slow down the switching speed because it would take a longer time to tune a wavelength over a wider range.

The architecture of a packet switch with limited-range wavelength converters does not differ from the switches described previously [29]. Output buffers are realized as FDLs in which n packets can be stored simultaneously (n wavelengths). The wavelength of each packet cannot be converted to any of the n available wavelengths due to the limited range of the wavelength converters, so each packet is not able to access all available wavelengths in an output buffer.

The wavelength conversion capability is measured using the wavelength conversion degree. A wavelength converter with conversion degree d is able to convert a wavelength to any wavelength of its d higher wavelengths and any of its d lower wavelengths. When $d = n$, the limited-range wavelength conversion becomes the same as the full-range wavelength conversion. Simulations carried out in [29] for various types of data traffic showed that, when the wavelength conversion capability reaches a certain threshold, the performance improvement is marginal if more wavelength conversion capability is subsequently added.

4) KEOPS (KEys to Optical Packet Switching): In 1995, the European ATMOS (ATM Optical Switching) project was succeeded by the KEOPS (KEys to Optical Packet Switching) project in which the study of the packet-switched optical network layer has been extended [18], [19]. The KEOPS proposal defines a multigigabit-per-second interconnection platform for end-to-end packetized information transfer that supports any dedicated electronic routing protocols and native WDM optical transmission.

In KEOPS, the duration of the packets is fixed; the header and its attached payload are encoded on a single wavelength carrier. The header is encoded at a low fixed bit rate to allow the utilization of standard electronic processing. The payload duration is fixed, regardless of its content; the data volume is proportional to the user-defined bit rate, which may vary from 622 Mb/s to 10 Gb/s, with easy upgrade capability. The fixed packet duration ensures that the same switch node can switch packets with variable bit rates. Consequently, the optical packet network layer proposed in KEOPS can be considered both bit rate and, to some degree, also transfer mode transparent, e.g., both ATM cells and IP packets can be switched.

The final report on the KEOPS project [19] suggests a 14-B packet header. Of that, 8 B are dedicated to a two-level hierarchy of routing labels. Then, 3 B are reserved for functionalities such as identification of payload type, flow control information, packet numbering for sequence integrity preservation, and header error checking. A 1-B pointer field flags the position of the payload relative to the header. Finally, 2 B are dedicated to the header synchronization pattern.

Each node in the KEOPS network has the following subblocks:

- an input interface, defined as a "coarse-and-fast" synchronizer that aligns the incoming packets in real time against the switch master clock;
- a switching core that routes the packets to their proper destination, solves contention, and manages the introduction of dummy packets to keep the system running in the absence of useful payload;

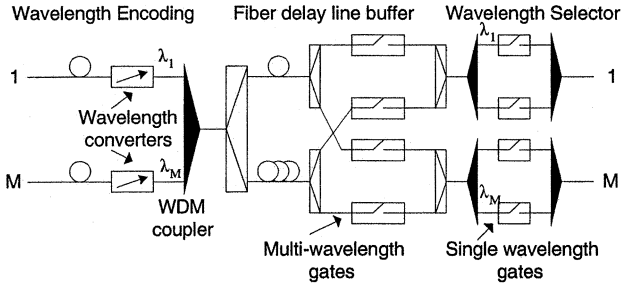

Fig. 15. Architecture of the broadcast and select switch suggested in KEOPS.

- an output interface that regenerates the data streams and provides the new header.

Two architectural options for the implementation of the switching fabric were evaluated exhaustively. The first one (wavelength routing switch) utilizes WDM to execute switching while the second one (broadcast and select switch) achieves high internal throughput due to WDM. Fig. 15 shows the broadcast and select switch suggested in KEOPS (electronic control not shown).

The principle of operation for the broadcast and select switch can be described as follows. Each incoming packet is assigned one wavelength through wavelength conversion identifying its input port and is then fed to the packet buffer. By passive splitting, all packets experience all possible delays. At the output of each delay line, multiwavelength gates select the packets belonging to the appropriate time slot. All wavelengths are gated simultaneously by these gates. Fast wavelength selectors are used to select only one of the packets, i.e., one wavelength. Multicasting can be achieved when the same wavelength is selected at more than one output. When the same wavelength is selected at all outputs, broadcasting is achieved.

5) The Data-Vortex Packet Switch: The data-vortex architecture [24], [30] was designed specifically to facilitate optical implementation by minimizing the number of the switching and logic operations and by eliminating the use of internal buffering. This architecture employs a hierarchical structure, synchronous packet clocking, and distributed-control signaling to avoid packet contention and reduce the necessary number of logic decisions required to route the data traffic. All packets within the switch fabric are assumed to have the same size and are aligned in timing when they arrive at the input ports. The timing and control algorithm of the switch permits only one packet to be processed at each node in a given clock frame, and therefore, the need to process contention resolution is eliminated. The wavelength domain is additionally used to enhance the throughput and to simplify the routing strategy.

The data-vortex topology consists of routing nodes that lie on a collection of concentric cylinders. The cylinders are characterized by a height parameter (H) corresponding to the number of nodes lying along the cylinder height, and an angle parameter (A), typically selected as a small odd number (< 10), corresponding to the number of nodes along the circumference. The total number of nodes is $(H)(A)$ for each of the concentric cylinders. The number of cylinders (C) scales with the height parameter as $C = \log_2(H)+1$. Because the maximum available number of input ports into the switch is given by $(H)(A)$, which equals the available number of output ports, the total number of routing nodes is given by $(H)(A)(\log_2 H + 1)$ for a switch fabric with input/output (I/O) ports.

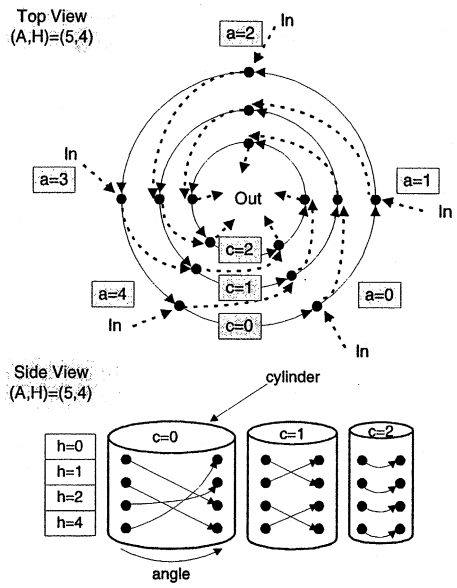

Fig. 16. Schematic of the data-vortex topology ($A = 5$, $H = 4$, $C = 3$).

In Fig. 16, an example of a switch fabric is shown. The routing tours are seen from the top and the side. Each cross point shown is the routing node, labeled uniquely by the coordinate (α, c, h), where $0 \leq \alpha < A$, $0 \leq c < C$, $0 \leq H < h$. Packets are injected at the outermost cylinder $(c = 0)$ from the input ports and emerge at the innermost cylinder $(c = \log_2 H)$ toward the output ports. Each packet is self-routed by proceeding along the angle dimension from the outer cylinder toward the inner cylinder. Every cylindrical progress fixes a specific bit within the binary header address. This hierarchical routing procedure allows the implementation of a technique of WDM-header encoding, by which the single-header-bit-based routing is accomplished by wavelength filtering at the header retrieval process. Since the header bits run at the rather low packet rate, there is no requirement of high-speed electronics within the node.

Packets are processed synchronously in a highly parallel manner. Within each time slot, every packet within the switch progresses by one angle forward in the given direction, either along the solid line toward the same cylinder or along the dashed line toward the inner cylinder. The solid routing pattern at the specific cylinder shown can be constructed as follows. First, we divide the total number of nodes along the height H into $2c$ subgroups, where c is the index of the cylinders. The first subgroup is then mapped as follows. For each step, we map half of the remaining nodes at angle (α) from the top to half of the remaining nodes at angle $(\alpha + 1)$ from bottom in a parallel way. This step is repeated until all nodes of the first subgroup are mapped from angle (α) to angle $(\alpha + 1)$. If multiple subgroups exist, the rest of them copy the mapping pattern of the first subgroup. The solid routing paths are repeated from angle to angle, which provide permutations between "1" and "0" for the specific header bit. At the same time, due to the smart twisting feature of the pattern, the packet-deflection probability is minimized because of the reduced correlation

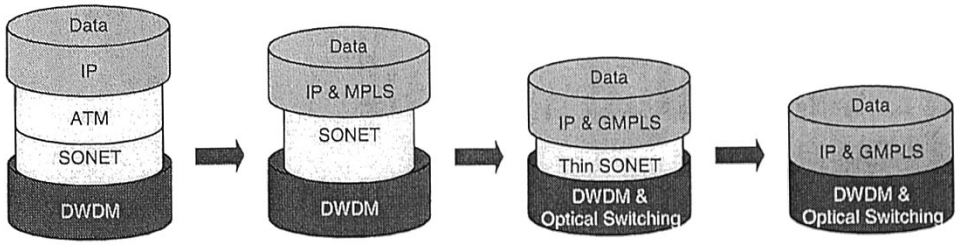

Fig. 17. The evolution toward photonic networking.

between different cylinders. The dashed-line paths between neighboring cylinders maintain the same height index h, because they are only used to forward the packets. By allowing the packet to circulate, the innermost cylinder also alleviates the overflow of the output-port buffers.

The avoidance of contention, and thereby reduction of processing necessary at the nodes, is accomplished with separate control bits. Control messages pass between nodes before the packet arrives at a given node to establish the right of way. Specifically, a node A on cylinder c has two input ports: one from a node B on the same cylinder c, and one from a node C on the outer cylinder $c-1$. A packet passing from B to A causes a control signal to be sent from B to C that blocks data at C from progressing to A. The blocked packet is deflected and remains on its current cylinder level. As mentioned previously, the routing paths along the angle dimension provide permutations between "1" and "0" for the specific header bit. Therefore, after two node hops, the packet will be in a position to drop to an inner cylinder and maintain its original target path. The control messages thus permit only one packet to enter a node in any given time period. Because the situation of two or more packets contending for the same output port never occurs in the data vortex, it significantly simplifies the logic operations at the node and, therefore, the switching time, contributing to the overall low latency of the switching fabric. Similar to convergence routing, the control mechanism and the routing topology of the data-vortex switch allow the packets to converge toward the destination after each routing stage. The fixed priority given to the packets at the inner cylinders by the control mechanism allows the routing fairness to be realized in a statistical sense. The data vortex has no internal buffers; however, the switch itself essentially acts as a delay-line buffer. Buffers are located at the input and output ports to control the data flow into and out of the switch. If there is congestion at an output buffer, the data waiting to leave to that buffer circulates around the lower cylinder and, thus, is optimally positioned to exit immediately as soon as the output ports are free.

IV. Generalized Multiprotocol Label Switching

Since 1995, there has been a dramatic increase in data traffic, driven primarily by the explosive growth of the Internet as well as the proliferation of virtual private networks, i.e., networks that simulate the operation of a private wide area network over the public Internet. As IP increasingly becomes the dominant protocol for data (and in the future voice and video) services, service providers and backbone builders are faced with a growing need to devise optimized network architectures for optical internetworks and the optical Internet [31]. The growth in IP traffic exceeds that of the IP-packet- processing capability. Therefore, the next-generation backbone networks should consist of IP routers with IP-packet-switching capability and OXCs with wavelength-path-switching capability to reduce the burden of heavy IP-packet-switching loads. This has raised a number of issues concerning the integration of the IP-routing functionality with the functionality offered by optical transport networks [32].

The outcome of this integration will enable service providers to carry a large volume of traffic in a cost-efficient manner and will thus improve the level of services provided. Current data network architectures do not seem capable of living up to the constantly increasing expectations [33]. Today's data networks typically have four layers: IP for carrying applications and services, ATM for traffic engineering, synchronous optical network/synchronous digital hierarchy (SONET/SDH) for transport, and dense wavelength-division multiplexing (DWDM) for capacity (Fig. 17). The IP layer provides the intelligence required to forward datagrams, while the ATM layer switches provide high-speed connectivity. Because there are two distinct architectures (IP and ATM), separate topologies, address spaces, routing and signaling protocols, as well as resource allocation schemes, have to be defined [33].

This architecture has been slow to scale for very large volumes of traffic and, at the same time, fairly cost-ineffective. Effective transport should optimize the cost of data multiplexing as well as data switching over a wide range of traffic volumes. It seems certain that DWDM and OXCs will be the preferred options for the transport and switching of data streams, respectively. Slower data streams will have to be aggregated into larger ones that are more suitable for DWDM and OXCs. In order to eliminate the SONET/SDH and ATM layers, their functions must move directly to the routers, OXCs, and DWDMs [33].

A. Multiprotocol Label Switching

Multiprotocol label switching (MPLS) [34]–[36] is a technique that attempts to bridge the photonic layer with the IP layer in order to allow for interoperable and scalable parallel growth in the IP and photonic dimensions. In a network that uses MPLS, all forwarding decisions are based on labels previously assigned to data packets. These labels are fixed-length values that are carried in the packets' headers. These values specify only the next hop and are not derived from other information contained in the header. Routers in an MPLS network are called label-switching routers (LSRs). Packets are forwarded from one LSR to another, thus forming label-switched paths (LSPs).

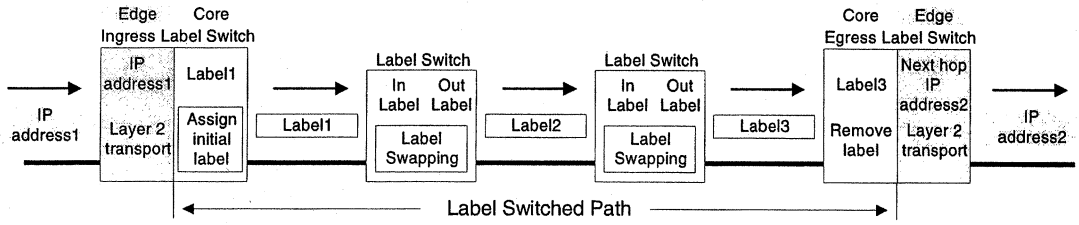

Fig. 18. Packet traversing a label switched path.

Labels are significant only in the current link and are used to identify a forwarding equivalence class (FEC). An FEC is a set of packets that are forwarded over the same path through a network. FECs are mapped to LSPs. Packets belonging to the same FEC do not necessarily have the same destination. Packets can be assigned to FECs, depending on their source and destination, QoS requirements, and other parameters. This is particularly advantageous for service providers. The network is so flexible, that new services can be added by simply modifying the way in which packets are assigned to FECs.

The separation of forwarding information from the content of the IP header allows the MPLS to be used with such devices as OXCs, whose data plane cannot recognize the IP header. LSRs forward data using the label carried by the data. This label, combined with the port on which the data was received, is used to determine the output port and outgoing label for the data.

The MPLS control component is completely separate from the forwarding component. The control component uses standard routing protocols to exchange information with other routers to build and maintain a forwarding table. When packets arrive, the forwarding component searches the forwarding table maintained by the control component to make a routing decision for each packet. Specifically, the forwarding component examines information contained in the packet's header, searches the forwarding table for a match, and directs the packet from the input interface to the output interface across the system's switching fabric. By completely separating the control component from the forwarding component, each component can be independently developed and modified. The only requirement is that the control component continues to communicate with the forwarding component by managing the packet-forwarding table [39].

The MPLS forwarding component is based on a label-swapping forwarding algorithm. The fundamental operations to this algorithm are the label distribution and the signaling operations. At the ingress nodes of the network, packets are classified and assigned their initial labels. In the core of the network, label switches ignore the packet's network layer header and simply forward the packet using the label-swapping algorithm. When a labeled packet arrives at a switch, the forwarding component uses the input port number and label to perform an exact match search of its forwarding table. When a match is found, the forwarding component retrieves the outgoing label, the outgoing interface, and the next-hop address from the forwarding table. The forwarding component then swaps the incoming label with the outgoing label and directs the packet to the outbound interface for transmission to the next hop in the LSP. When the labeled packet arrives at the egress label switch, the forwarding component searches its forwarding table. If the next hop is not a label switch, the egress switch discards the label and forwards the packet using conventional longest-match IP forwarding. Fig. 18 illustrates the course of a packet traversing an LSP [39].

Label swapping provides a significant number of operational benefits when compared with conventional hop-by-hop network layer routing. These benefits include tremendous flexibility in the way that packets are assigned to FECs, the ability to construct customized LSPs that meet specific application requirements, and the ability to explicitly determine the path that the traffic will take across the network.

The MPLS framework includes significant applications, such as constraint-based routing. This allows nodes to exchange information about network topology, resource availability, and even policy information. This information is used by the algorithms that determine paths to compute paths subject to specified resource and/or policy constraints. After the computation of paths, a signaling protocol such as the Resource Reservation Setup Protocol (RSVP) is used to establish the routes that were computed, and thus, the LSP is created. Then, the MPLS data plane is used to forward the data along the established LSPs. Constraint-based routing is used today for two main purposes: traffic engineering (replacement for the ATM as the mechanism for traffic engineering) and fast reroute (an alternative to SONET as a mechanism for protection/restoration). In other words, enhancements provided by the MPLS to IP routing make it possible to bypass ATM and SONET/SDH by migrating functions provided by these technologies to the IP/MPLS control plane.

B. Generalized Multiprotocol Label Switching

GMPLS extends MPLS to support not only devices that perform packet switching, but also those that perform switching in the time, wavelength, and space domains. This requires modifications to current signaling and routing protocols and has also triggered the development of new protocols, such as the link management protocol (LMP) [40]. Wavelength paths, called optical LSPs (OLSPs), are set and released in a distributed manner based on the functions offered by the GMPLS. LSRs are able to dynamically request bandwidth from the optical transport network. The establishment of the necessary connections is handled by OXCs, which use labels that map to wavelengths, fibers, time slots, etc. This implies that the two control planes (for LSRs and OXCs) are in full cooperation. However, GMPLS does not specify whether these two control planes are integrated or separated.

This means that there can be two operational models as well as a hybrid approach that combines these two models. The overlay model [33] hides details of the internal network, resulting in two separate control planes with minimal interaction between them. One control plane operates within the core optical network, and the other between the core and the surrounding edge devices [this control plane is called the user network interface (UNI)]. The edge devices support lightpaths that are either dynamically signaled across the core optical network or statically provisioned without seeing inside the core's topology. On the contrary, according to the peer model [33], edge devices are able to see the topology of the network. This allows edge devices to participate in the making of routing decisions, thus eliminating the artificial barriers between the transport and the routing domains. According to the hybrid model, some devices function as peers, thus implementing an integrated control plane, while others have their separate control planes and interface with the core network through the UNI.

It is evident that, in order to employ GMPLS, OXCs must be able to exchange control information. One way to support this is to preconfigure a dedicated control wavelength between each pair of adjacent OXCs, or between an OXC and a router, and to use this wavelength as a supervisory channel for exchange of control traffic. Another possibility is to construct a dedicated out-of-band IP network for the distribution of control traffic.

Consider a network as a directed graph whose nodes are network elements (MPLS switches, cross-connects, etc.) and whose edges are links (fibers, cables, etc.). Each edge in the graph has associated attributes, such as IP addresses, cost, and unreserved bandwidth. A link-state protocol allows all the nodes to dynamically coordinate a coherent up-to-date picture of this graph, including the attributes of each edge. This picture of the graph is referred to as the link-state database. Once the link-state database is synchronized among all participating routers, each router uses the database to construct its own forwarding table. When a packet arrives at a router, the forwarding table is then consulted to determine how to forward the packet. Should the status of any link be changed, including adding or removing links, the link-state database must be resynchronized, and all of the routers must recalculate their forwarding tables using the updated information in the link-state database.

Several enhancements to GMPLS have been proposed. In [38], an overview of signaling enhancements and recovery techniques is presented, while [37] concerns itself with the challenges and possible enhancements regarding optical network restoration. In [33], enhancements to routing and management are suggested in order to address the following issues.

1) MPLS LSPs can be allocated bandwidth from a continuous spectrum, whereas optical/TDM (time-division-multiplexed) bandwidth allocation is from a small discrete set of values.
2) Today, there are rarely more than ten parallel links between a pair of nodes. To handle the growth of traffic, providers will need to deploy hundreds of parallel fibers, each carrying hundreds of lambdas between a pair of network elements. This, in turn, raises three subissues.
    a) The overall number of links in an optical/TDM network can be several orders of magnitude larger than that of an MPLS network.
    b) Assigning IP addresses to each link in an MPLS network is not particularly onerous; assigning IP addresses to each fiber, lambda, and TDM channel is a serious concern because of both the scarcity of IP addresses and the management burden.
    c) Identifying which port on a network element is connected to which port on a neighboring network element is also a major management burden and highly error-prone.
3) Fast fault detection and isolation and fast failover to an alternate channel are needed.
4) The user data carried in the optical domain is transparently switched to increase the efficiency of the network. This necessitates transmitting control-plane information decoupled from user data.

Some of the enhancements that aim to resolve the issues mentioned here follow.

Link Bundling: As mentioned previously, the link-state database consists of all the nodes and links in a network, along with the attributes of each link. Because of the large number of nodes, the link-state database for an optical network can easily be several orders of magnitude bigger than that for an MPLS network.

To address this issue, several parallel links of similar characteristics can be aggregated to form a single "bundled" link [33]. This reduces the size of the link-state database by a large factor and improves the scalability of the link-state protocol. By summarizing the attributes of several links into one bundled link, some information is lost; for example, with a bundle of SONET links, the switching capability of the link interfaces are flooded.

Unnumbered Links: All the links in an MPLS network are typically assigned IP addresses. When a path is computed through the network, the links that constitute the path are identified by their IP addresses; this information is conveyed to the signaling protocol, which then sets up the path. Thus, it would seem that every link must have an IP address. However, this is very difficult for optical networks, because of their large number of links.

This problem can be solved if each network node numbers its links internally [33]. Each node is identified by a unique router ID. In order to identify a particular link, the tuple [routerID, link number] is used. The reduction of management effort in configuring IP addresses, tracking allocated IP addresses, and dealing with the occasional duplicate address allocation is a significant savings, especially in the context of optical networks.

Link Management Protocol: The LMP runs between adjacent nodes and is used for both link provisioning and fault isolation [33]. A key service provided by LMP is the associations between neighboring nodes for the component link IDs that may, in turn, be used as labels for physical resources. These associations do not have to be configured manually, a poten-

tially error-prone process. A significant improvement in manageability accrues because the associations are created by the protocol itself.

Within a bundled link, the component links and associated control channel need not be transmitted over the same physical medium. LMP allows for decoupling of the control channel from the component links. For example, the control channel could be transmitted along a separate wavelength or fiber or over a separate Ethernet link between the two nodes. A consequence of allowing the control channel for a link to be physically diverse from the component links is that the health of a control channel of a link does not correlate to the health of the component links, and vice versa. Furthermore, due to the transparent nature of photonic switches, traditional methods can no longer be used to monitor and manage links.

LMP is designed to provide four basic functions for a node pair: control channel management, link connectivity verification, link property correlation, and fault isolation. Control channel management is used to establish and maintain connectivity between adjacent nodes. The link verification procedure is used to verify the physical connectivity of the component links. The LinkSummary message of LMP provides the correlation function of link properties (e.g., link ID's, protection mechanisms, and priorities) between adjacent nodes. This is done when a link is first brought up and may be repeated any time a link is up and not in the verification procedure. Finally, LMP provides a mechanism to isolate link and channel failures in both opaque and transparent networks, independent of the data format.

In [32], an Open Shortest Path First (OSPF) extension is proposed for the realization of multilayer traffic engineering. According to this extension, each node advertizes both the number of total wavelengths and the number of unused wavelengths for each link. This information is passed to all egress nodes using an extended signaling protocol, so that they are aware of the link state, and it is taken into account in the establishment of new optical LSPs. Additional extensions aim at the minimization of the number of wavelength conversions needed. Wavelength conversion is a very expensive operation in all-optical photonic networks. Apart from the protocol extensions, a heuristics-based multiplayer topology design scheme is proposed in [32]. This scheme uses IP traffic measurements in a GMPLS switch to yield the OLSP that minimizes network cost, in response to fluctuations in IP traffic demand. In other words, the OLSP network topology is dynamically reconfigured to match IP traffic demand.

From a service provider's perspective, GMPLS offers advanced management and control. Carriers will greatly benefit from a single common set of management semantics that unifies heterogeneous optical networks and delivers consistent information across all elements. The lack of such unified management information throttles today's optical networks, limiting performance and cost-effectiveness [40].

C. Automatically Switched Optical Network

The ITU-T recommendation G.8080/Y.1304 specifies the architecture and requirements for the automatically switched optical network (ASON) [41]. This recommendation describes the set of control-plane components that are used to manipulate transport network resources in order to provide the functionality of setting up, maintaining, and releasing connections. The ASON is meant to serve as a reference architecture for service providers and protocol developers. This section provides an overview of the main features of the ASON as well as some of the differences with respect to the GMPLS.

The purpose of the ASON control plane is to

- facilitate fast and efficient configuration of connections within a transport layer network to support both switched (user requests) and soft permanent connections (management request);
- reconfigure or modify connections that support calls that have previously been set up;
- perform a restoration function.

The ASON control plane will be subdivided into domains that match the administrative domains of the network. Within an administrative domain, the control plane may be further subdivided, e.g., by actions from the management plane. This allows the separation of resources into, for example, domains for geographic regions, that can be further divided into domains that contain different types of equipment. Within each domain, the control plane may be further subdivided into routing areas for scalability, which may also be further subdivided into sets of control components. The transport-plane resources used by the ASON will be partitioned to match the subdivisions created within the control plane [41].

The interconnection between domains, routing areas, and, where required, sets of control components is described in terms of reference points. The reference point between an administrative domain and an end user is the UNI. The reference point between domains is the external network–network interface (E-NNI). The reference point within a domain between routing areas and, where required, between sets of control components within routing areas is the internal network–network interface (I-NNI). Each reference point has different requirements on the degree of information hiding [42]. In particular, the UNI hides all routing and addressing information pertaining to the interior of the network from the user. The ASON is very clear on the fact that users should belong to a different address space from internal network nodes. The I-NNI is a trusted reference point. Full routing information can be flooded. The E-NNI lies somewhere in between.

The ASON control plane separates call control from connection control. Call control is a signaling association between one or more user applications and the network to control the setup, release, modification, and maintenance of sets of connections. Call control is used to maintain the association between parties, and a call may embody any number of underlying connections, including zero, at any instance of time. In order to establish a call, signaling messages are exchanged between the calling and the called parties and the network. Specifically, the calling party's call controller contacts the network call controller, which, in turn, contacts the call controller of the called party. After a call has been accepted, it may request the establishment of one or more connections. Decisions concerning the

establishment of connections are made by the connection controller. It must be noted that the connection controller communicates only with the network call controller and not with any of the parties involved in the call.

The information needed to determine whether a connection can be established is provided by the routing controller. There are two roles of the routing controller.

- The routing controller responds to requests from connection controllers for path (route) information needed to set up connections. This information can vary from end-to-end (e.g., source routing) to next hop.
- The routing controller responds to requests for topology information for network management purposes.

If the requirements for a specific connection are not met, the parties involved may renegotiate the connection parameters without terminating the current call. When a call needs to be terminated, signaling messages must also be exchanged.

An obvious difference between the ASON and the GMPLS is the way in which the network is envisaged. GMPLS switches are seen as operating in a GMPLS-only cloud of peer network elements. Nodes at the edge of the cloud are capable of accepting non-GMPLS protocol data and tunneling it across the GMPLS cloud to other edge nodes. All the nodes and links that constitute the GMPLS network share the same IP address space, and the GMPLS implies a trusted environment. On the contrary, the ASON views the network as one composed of domains that interact with other domains in a standardized way, but whose internal operation is protocol independent and not subject to standardization [42].

In the ASON, the distinction between the users and the network is clear. This implies that, in contrast with the GMPLS, new addresses need to be assigned to users of the network in order to maintain complete separation of the user and the network addressing spaces. Next, because no routing information is allowed to flow across the UNI, the users cannot calculate suitable routes on their own. Apart from that, in the GMPLS, a link is defined to be capable of supporting multiple different layers of switched traffic, while in the ASON, a link is defined to be capable of carrying only a single layer of switched traffic. This means, that the ASON treats each layer separately, i.e., there is a layer-specific instance of the signaling, routing, and discovery protocols running for each layer [42].

V. Optical Burst Switching

OBS is an attempt at a new synthesis of optical and electronic technologies that seeks to exploit the tremendous bandwidth of optical technology, while using electronics for management and control [43]. Burst switching is designed to facilitate switching of the user data channels entirely in the optical domain.

This approach divides the entire network in two regions: edge and core. At the edge, the usual packets are assembled with some procedures to form bursts, i.e., collections of packets that have certain features in common (e.g., destination). Bursts are assigned to wavelength channels and are switched through transparently without any conversion. Bursts travel only in the core nodes, and when they arrive at the network edge, they are disassembled into the original packets and delivered with the usual methods.

A burst-switched network that has been properly designed can be operated at reasonably high levels of utilization, with acceptably small probability that a burst is discarded due to lack of an available channel or storage location, thus achieving very good statistical multiplexing performance. When the number of wavelength channels is large, reasonably good statistical multiplexing performance can be obtained with no burst storage at all [49].

A. Network and Node Architecture

The transmission links in a burst-switched system carry multiple channels, any one of which can be dynamically assigned to a user data burst. The channels are wavelength-division multiplexed. One channel (at least) on each link is designated as a control channel, and is used to control dynamic assignment of the remaining channels to user data bursts. It is also possible to provide a separate fiber for each channel in a multifiber cable.

In principle, burst transmission works as follows. Arriving packets are assembled to form bursts at the edge of the OBS network. The assembly strategy is a key design issue and is discussed later in this paper. Shortly before the transmission of a burst, a control packet is sent in order to reserve the required transmission and switching resources. Data is sent almost immediately after the reservation request without receiving an acknowledgment of successful reservation. Although there is a possibility that bursts may be discarded due to lack of resources, this approach yields extremely low latency, since propagation delay usually dominates transmission time in wide area networks [44].

The reservation request (control packet) is sent on the dedicated wavelength some offset time prior to the transmission of the data burst. This basic offset has to be large enough to electronically process the control packet and set up the switching matrix for the data burst in all nodes. When a data burst arrives in a node, the switching matrix has been already set up, i.e., the burst is kept in the optical domain. The format of the data sent on the data channels is not constrained by the burst-switching system. Data bursts may be IP packets, a stream of ATM cells, or frame relay packets. However, since the burst-switching system must be able to interpret the information on the control channel, a standard format is required here [49]. The control packet includes a length field specifying the amount of data in the burst, as well as an offset field that specifies the time between the transmission of the first bit of the control packet and the first bit of the burst.

B. Burst Generation

The way in which packets are assembled to form bursts can heavily affect the network performance. The assembly method determines the network traffic characteristics. The process in which bursts are assembled must take several parameters into account, such as the destination of the packets or their QoS requirements.

There must be a minimum requirement on the burst size. A burst must be sufficiently long in order to allow the node receiving the control packet to convert it into an electronic form,

to elaborate and to update it (if necessary), and to prepare the switching fabric. This requirement is also dictated by the limited capacity of the control channel. If bursts are made too small, the corresponding control packets may exceed the capacity of the control channel. Conversely, if data traffic is not intense, the burst generator must not delay bursts indefinitely until the minimum size requirement is met. After a specified time period has elapsed, existing packets must be joined to form a burst, and, if necessary, the burst will be padded to achieve a minimum length.

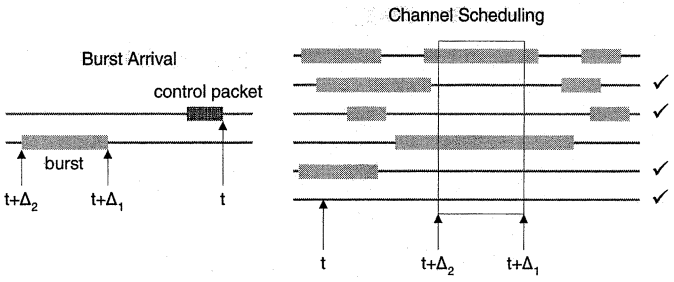

Fig. 19. Horizon channel scheduling.

C. Channel Scheduling

In order to efficiently handle bursts that may be as short as 1 μs in duration, the resource management mechanisms in a burst switch must have the ability to project resource availability into the future and make decisions based on future, rather than current, resource availability. This type of resource management is referred to as look-ahead resource management [49].

Suppose that a control packet is received and that the burst will arrive in 10 μs. If an available channel is assigned to the burst, the resources of that channel are fragmented, leaving a 10-μs idle period, which may be difficult to use. It would be preferable to assign the burst to a channel that will become available just shortly before the burst arrives. Apart from that, the channel scheduling algorithm must be kept as simple as possible in order to support high burst processing rates.

One simple technique for channel assignment is horizon scheduling. In horizon scheduling, the controller for a link maintains a time horizon for each of the channels of an outgoing link. The horizon is defined as the earliest time after which there is no planned use of the channel. The horizon scheduler assigns arriving bursts to the channel with the latest horizon that is earlier than the arrival time of the burst, if there is such a channel. If there is no such channel, the burst is assigned to the channel with the smallest horizon and is diverted to the burst storage area, where it is delayed until the assigned channel is available [43]. Once a channel has been selected, the scheduling horizon is recomputed to be equal to the time when the burst is due to be completed (based on knowledge of the arrival time and burst duration).

Fig. 19 illustrates an example of horizon scheduling. The checkmarks denote the available channels.

Horizon scheduling is straightforward to implement in hardware, but because it does not keep track of time periods before a channel's horizon when the channel is unused, it cannot insert bursts into these open spaces. The horizon scheduler can provide good performance if the time between the arrival of a control packet and the subsequent arrival of a burst is subject to only small variations. However, if the variations are as large or larger than the time duration of the bursts, the performance of horizon scheduling can deteriorate significantly.

An improvement can be made on the efficiency of horizon scheduling by processing bursts out of order. Rather than process bursts as soon as their control packets arrive, one can delay processing and then process the bursts in the order of expected burst arrival, rather than the order in which the control packets arrived. As the control packets arrive, they are inserted into a resequencing buffer, in the order in which the bursts are to arrive. A horizon scheduler then processes requests from the resequencing buffer. The processing of a request is delayed until shortly before the burst is to arrive, reducing the probability that we later receive a control packet for a burst that will arrive before any of the bursts that have already been scheduled. If the lead time for processing bursts is smaller than the burst durations, optimal performance can be obtained.

A more sophisticated version of the horizon scheduling algorithm tries to fill the voids in the scheduling of bursts in order to fully exploit the available bandwidth. This is accomplished by searching for the channel that has the shortest gap that is closest to the burst's arrival and can accommodate the burst. This algorithm exploits the available bandwidth efficiently, but its increased complexity may result in an increase in the burst discard probability.

D. QoS Support

Burst-switching systems are capable of providing support for differentiated classes of service. As it was mentioned previously, the burst assembly mechanism can take into account a packet's class of service and thus form bursts based on such criteria. In order for the burst switch to distinguish between bursts of different classes, additional information could be placed in the control packets. This, however, is not desirable, since it increases the processing overhead and the complexity of the link scheduling algorithm. Another thought would be to associate classes of service with WDM channels. This approach, however, could result in a waste of bandwidth.

The preferred way in which multiple classes of service are implemented in a burst-switched system is by manipulating the offset time between the control packet and the burst. The bigger the offset of the control packet from the burst, the more time the switch has to prepare for the arrival of the burst. Since the switch is aware of the arrival of the burst a long time before it actually arrives, there is a high probability of finding a free channel and thus a lower probability that the burst will be discarded.

It is true that bursts belonging to high-priority service classes will have to wait longer before they are scheduled. However, this delay is experienced only at the edge of the network and can be tolerated by most applications. Fig. 20 illustrates a scenario with three wavelengths in which a high-priority and a low-priority burst arrive at the same time. It can be seen that the low-priority burst cannot be served, since all wavelengths are already occupied during its transmission time, whereas the high priority burst is able to find a wavelength due to its much larger offset [44].

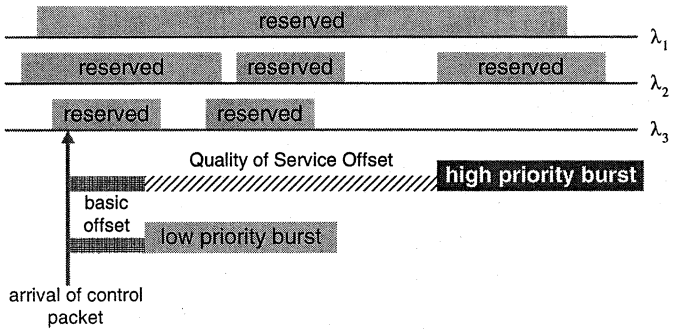

Fig. 20. Reservation scenario for bursts of different classes.

E. Contention Resolution

One of the key problems in the application of burst switching is the handling of burst contentions that take place when two or more incoming bursts are directed to the same output line [45]. While the contention resolution techniques that were described previously (buffering, deflection routing, and wavelength conversion) can be applied to burst-switched networks, additional schemes may also be necessary in order to increase the network throughput and utilization.

When traditional contention resolution schemes are applied, an entire burst is discarded when the contention cannot be resolved, even though the overlap between the contending bursts may be minimal. Instead of dropping the entire burst (and thus several packets), optical composite burst switching (OCBS) dictates that only the initial part of the burst is to be discarded, until a wavelength channel becomes available on the output fiber. From that instant, the switch will transmit the remainder of the truncated burst [46]. OCBS attempts to minimize the packet loss probability, rather than the burst loss probability, and thus allows the switch throughput to be increased in terms of the number of accepted packets. According to an analytical model presented in [46], OCBS supports significantly more traffic than OBS for a given packet loss probability. The significant improvement achieved by OCBS is due to the fact that, on average, the part of the truncated burst that is lost is significantly smaller than its successfully transmitted part.

The authors of [47] suggest a contention resolution technique called burst segmentation. According to this technique, each burst is divided into basic transport units called segments. Each of these segments may consist of a single packet or multiple packets. The segments define the possible partitioning points of a burst when the burst is in the optical network. When contention occurs, only those segments of a given burst that overlap with segments of another burst will be dropped. When two bursts are in contention, either the tail of the first burst or the head of the second burst will be truncated.

Burst segmentation can be combined with other methods for contention resolution, such as deflection. The segments that would otherwise be discarded are deflected to an alternate port. Implementing segmentation with deflection increases the probability of the burst reaching the destination and, hence, improves the performance [47]. The authors of [48] show how segmentation with deflection can be used to provide differentiated services in OBS networks. In such a scheme, the priorities

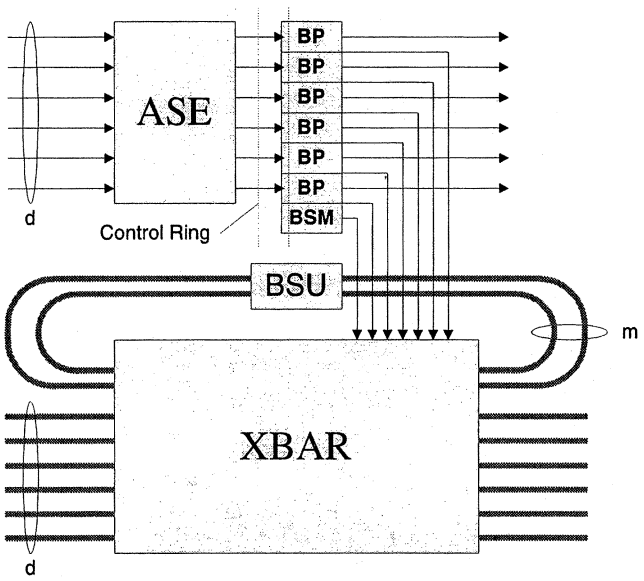

Fig. 21. Burst switch element.

of the contending bursts, as well as other factors, such as the lengths of the overlapping segments, are taken into account when deciding which burst will be segmented and/or deflected.

F. "Terabit Burst Switching"

Washington University's Terabit Burst Switching Project [49] seeks to demonstrate the feasibility of OBS. The project will lead to the construction of a demonstration switch with throughput exceeding 200 Gb/s and scalable to over 10 Tb/s.

The suggested burst switch architecture consists of a set of *input/output modules* (IOM) that interface to external links and a multistage interconnection network of *burst switch elements* (BSE). Control packets are referred to as burst header cells (BHCs). The IOM uses the address information contained in BHCs to do a routing table lookup. The result of this lookup includes the number of the output port to which the burst is to be forwarded. This information is inserted into the BHC, which is then forwarded to the first-stage BSE. The data channels pass directly through the IOMs but are delayed at the input by a fixed interval to allow time for the control operations to be performed.

When a BHC is passed to a BSE, the control section of the BSE uses the output port number in the BHC to determine which of its output links to use when forwarding the burst. If the required output link has an idle channel available, the burst is switched directly through to that output link. If no channel is available, the burst can be stored within a shared *burst storage unit* (BSU) within the BSE.

Each BSE (Fig. 21) in a burst switch requires a wavelength converting switch, capable of switching a signal from any input of the BSE's d input fibers to any of its d output fibers. The control section consists of a d port ATM switch element (ASE), a set of d burst processors (BP), and a burst storage manager (BSM). The data path consists of a crossbar switch, together with the burst storage unit (BSU). The BSU is connected to the crossbar with m input links and m output links. Each BP is

responsible for handling bursts addressed to a particular output link. When a BP is unable to switch an arriving burst to a channel within its output link, it requests use of one of the BSU's storage locations from the BSM, which switches the arriving burst to an available storage location (if there is one). Communication between the BSEs and the BSM occurs through a local control ring provided for this purpose.

To enable effective control of dynamic routing, BSEs provide flow control information to their upstream neighbors in the interconnection network. The burst processor receives status information from downstream neighbors and other BPs in the same BSE (through the control ring) and updates the stored status information accordingly. It also forwards status information to one of the neighboring upstream BSEs and the other BPs in the same BSE.

VI. Conclusion

Existing long-haul networks seem unable to meet market demands for network capacity. The capacity of the transmission medium (optical fiber) is not fully exploited. In order to exploit the tremendous capacity provided by optical fiber, the switching functions must, even partially, be executed optically.

The first step toward the migration of the switching function from electronics to optics is the replacement of the switching core with an all-optical switch fabric. The major technologies for all-optical switching were presented in this paper. Although, there are several commercially available all-optical switches, carriers appear reluctant to replace the switching cores of their networks. That, of course, is understandable, considering that there is a huge financial investment involved. Electronic switching has served everyone well for a number of years, and optical switching is by no means a mature technology.

The next step toward all-optical switching is the deployment of a switching technique specifically designed for optical networks. All three of the optical switching techniques that were presented face significant challenges. The lack of optical memory as well as the lack of processing capabilities in the optical domain seem to be the greatest obstacles. Researchers are looking for ways to tackle these obstacles and provide optical networks with the flexibility and efficiency that everyone needs.

Major technical difficulties will need to be overcome [50]–[52] on the way toward photonic networking. Nevertheless, one must always keep in mind that future scientific breakthroughs may counteract the fundamental limitations of optics and thus completely change the current outlook of networking.

References

[1] R. Ramaswami and K. N. Sivarajan, *Optical Networks, A Practical Perspective*. San Fransisco, CA: Morgan Kaufmann, 1998.

[2] P. B. Chu, S.-S. Lee, and S. Park, "MEMS: The path to large optical cross-connects," *IEEE Commun. Mag.*, pp. 80–87, Mar. 2002.

[3] D. J. Bishop, C. R. Giles, and G. P. Austin, "The Lucent lambdarouter: MEMS technology of the future here today," *IEEE Commun. Mag.*, pp. 75–79, Mar. 2002.

[4] P. De Dobbelaere, K. Falta, L. Fan, S. Gloeckner, and S. Patra, "Digital MEMS for optical switching," *IEEE Commun. Mag.*, pp. 88–95, Mar. 2002.

[5] A. Dugan, L. Lightworks, and J.-C. Chiao, "The optical switching spectrum: A primer on wavelength switching technologies," *Telecommun. Mag.*, May 2001.

[6] S. Bregni, G. Guerra, and A. Pattavina, "State of the art of optical switching technology for all-optical networks," in *Communications World*. Rethymo, Greece: WSES Press, 2001.

[7] K. Sakuma, H. Ogawa, D. Fujita, and H. Hosoya, "Polymer Y-branching thermo-optic switch for optical fiber communication systems," in The 8th Microoptics Conf. (MOC'01), Osaka, Japan, Oct. 24–26, 2001.

[8] J.-C Chiao, "Liquid crystal optical switches," in 2001 OSA Topical Meetings, Photonics in Switching, Monterey, CA, June 11–15, 2001.

[9] J. Sapriel, D. Charissoux, V. Voloshinov, and V. Molchanov, "Tunable acoustooptic filters and equalizers for WDM applications," *J. Lightwave Technol.*, vol. 20, pp. 892–899, May 2002.

[10] T. E. Stern and K. Bala, *Multiwavelength Optical Networks, A Layered Approach*. Reading, MA: Addison-Wesley, 1999.

[11] J. E. Fouquet *et al.*, "A compact, scalable cross-connect switch using total internal reflection due to thermally-generated bubbles," in *IEEE LEOS Annu. Meeting*, Orlando, FL, 1988, pp. 169–170.

[12] B. Mukherjee, *Optical Communication Networks*. New York: McGraw-Hill, 1997.

[13] V. E. Beneš, *Mathematical Theory of Connecting Networks*. New York: Academic, 1965.

[14] D. K. Hunter and I. Andonovic, "Approaches to optical internet packet switching," *IEEE Commun. Mag.*, pp. 116–122, Sept. 2000.

[15] V. Eramo and M. Listanti, "Packet loss in a bufferless optical WDM switch employing shared tunable wavelength converters," *J. Lightwave Technol.*, vol. 18, pp. 1818–1833, Dec. 2000.

[16] S. Yao, B. Mukherjee, and S. Dixit, "Advances in photonic packet switching: An overview," *IEEE Commun. Mag.*, pp. 84–94 Feb., 2000.

[17] P. B. Hansen, S. L. Danielsen, and K. E. Stubkjaer, "Optical packet switching without packet alignment," in 24th European Conf. Optical Communications, Madrid, Spain, Sept. 1998.

[18] S. L. Danielsen, P. B. Hansen, and K. E. Stubkjaer, "Wavelength conversion in optical packet switching," *J. Lightwave Technol.*, vol. 16, pp. 2095–2108, Dec. 1998.

[19] "KEOPS: KEys to Optical Packet Switching, Final <ReportRef. 19 - December Ref. 22 - Member, IEEE Ref. 24 - Student Member, IEEE and Member, IEEE, Member, OSA Ref. 28 - Member, IEEE Ref. 29 - Member, IEEE and Senior Member, IEEE and Student Member, IEEE Ref. 30 - Student Member, IEEE and Member, IEEE Ref. 31 - Member, IEEE and Associate Member, IEEE Ref. 32 - NTT Corporation Ref. 33 - Calient Networks and Juniper Networks and Cisco Systems Ref. 39 - Marketing Engineer, Juniper Networks 2. We have tagged Refs. 53-9 doctype as other.," ACTS Project, AC043, 1998.

[20] M. Murata and K. Kitayama, "Ultrafast photonic label switch for asynchronous packets of variable length," in *Proc*. IEEE INFOCOM 2002, New York, June 23–27, 2002.

[21] I. Glesk, K. I. Kang, and P. R. Prucnal, "Ultrafast photonic packet switching with optical control," *Opt. Express*, vol. 1, no. 5, pp. 126–132, Sept. 1997.

[22] S. L. Danielsen, C. Joergensen, B. Mikkelsen, and K. E. Stubkjaer, "Analysis of a WDM packet switch with improved performance under bursty traffic conditions due to tunable wavelength converters," *J. Lightwave Technol.*, vol. 16, pp. 729–735, May 1998.

[23] A. Ge, L. Tancevski, G. Castanon, and L. S. Tamil, "WDM fiber delay line buffer control for optical packet switching," in *Proc. SPIE OptiComm 2000*, Oct. 2000, pp. 1671–1673.

[24] Q. Yang, K. Bergman, G. D. Hughes, and F. G. Johnson, "WDM packet routing for high-capacity data networks," *J. Lightwave Technol.*, vol. 19, no. 10, pp. 1420–1426, Oct. 2001.

[25] V. Eramo and M. Listanti, "Packet loss in a bufferless optical WDM switch employing shared tunable wavelength converters," *J. Lightwave Technol.*, vol. 18, pp. 1818–1833, Dec. 2000.

[26] J. P. Jue, "An algorithm for loopless deflection in photonic packet-switched networks," in *Proc. IEEE ICC '02*, vol. 5, New York, NY, Apr. 2002, pp. 2776–2780.

[27] S. L. Danielsen, C. Joergensen, B. Mikkelsen, and K. E. Stubkjaer, "Optical packet switched network layer without optical buffers," *IEEE Photon. Technol. Lett.*, vol. 10, pp. 896–898, June 1998.

[28] S. L. Danielsen, B. Mikkelsen, C. Joergensen, T. Durhuus, and K. E. Stubkjaer, "WDM packet switch architectures and analysis of the influence of tuneable wavelength converters on the performance," *J. Lightwave Technol.*, vol. 15, pp. 219–227, Feb. 1997.

[29] G. Shen, S. K. Bose, T. H. Cheng, C. Lu, and T. Y. Chai, "Performance study on a WDM packet switch with limited-range wavelength converters," *IEEE Commun. Lett.*, vol. 5, pp. 432–434, Oct. 2001.

[30] Q. Yang and K. Bergman, "Traffic control and WDM routing in the data vortex packet switch," *IEEE Photon. Technol. Lett.*, vol. 14, no. 2, pp. 236–238, Feb. 2002.
[31] A. Rodriguez-Moral, P. Bonenfant, S. Baroni, and R. Wu, "Optical data networking: Protocols, technologies, and architectures for next generation optical transport networks and optical internetworks," *J. Lightwave Technol.*, vol. 18, pp. 1855–1870, Dec. 2000.
[32] K.-I. Sato, N. Yamanaka, Y. Takigawa, M. Koga, S. Okamoto, K. Shiomoto, E. Oki, and W. Imajuku, "GMPLS-Based photonic multilayer router (Hikari router) architecture: An overview of traffic engineering and signaling technology," *IEEE Commun. Mag.*, pp. 96–101, Mar. 2002.
[33] A. Banerjee, J. Drake, J. P. Lang, B. Turner, K. Kompella, and Y. Rekhter, "Generalized multiprotocol label switching: An overview of routing and management enhancements," *IEEE Commun. Mag.*, pp. 144–150, Jan. 2001.
[34] G. Armitage, "MPLS: The magic behind the myths," *IEEE Commun. Mag.*, pp. 124–131, Jan. 2000.
[35] T. Li, "MPLS and the evolving internet architecture," *IEEE Commun. Mag.*, pp. 38–41, Dec. 1999.
[36] D. O. Awduche, "MPLS and traffic engineering in IP networks," *IEEE Commun. Mag.*, pp. 42–47, Dec. 1999.
[37] R. Doverspike and J. Yates, "Challenges for MPLS in optical network restoration," *IEEE Commun. Mag.*, pp. 89–96, Feb. 2001.
[38] A. Banerjee, J. Drake, J. Lang, B. Turner, D. Awduche, L. Berger, K. Kompella, and Y. Rekhter, "Generalized multiprotocol label switching: An overview of signaling enhancements and recovery techniques," *IEEE Commun. Mag.*, pp. 144–151, July 2001.
[39] C. Semeria. Multiprotocol Label Switching Enhancing Routing in the New Public Network. White Paper. Juniper Networks. [Online]. Available: http://www.juniper.net/techcenter/techpapers/200001.html
[40] L. Ceuppens. GMPLS: Approaching Implementation in Photonic Networks. Calient Networks. [Online]. Available: www.mplsworld.com/archi_drafts/focus/analy-ceuppens.htm
[41] *Architecture for the Automatically Switched Optical Network (ASON)*, ITU-T Recommendation G.ason (G.8080/Y.1304), Oct. 2001.
[42] N. Larkin. ASON and GMPLS—The Battle of the Optical Control Plane. White Paper. [Online]. Available: http://www.dataconnection.com/mpls/#whitepapers
[43] J. S. Turner, "WDM burst switching for petabit data networks," in *Proc. Optical Fiber Conf.*, Mar. 2000.
[44] K. Dolzer and C. Gauger, "On burst assembly in optical burst switching networks—A performance evaluation of just-enough-time," in 17th Int. Teletraffic Congr., Salvador, Brazil, Sept. 24–28, 2001.
[45] A. Detti, V. Eramo, and M. Listanti, "Performance evaluation of a new technique for IP support in a WDM optical network: Optical composite burst switching (OCBS)," *J. Lightwave Technol.*, vol. 20, pp. 154–165, Feb. 2002.
[46] M. Neuts, Z. Rosberg, H. L. Vu, J. White, and M. Zukerman, "Performance analysis of optical composite burst switching," *IEEE Commun. Lett.*, vol. 6, pp. 346–348, Aug. 2002.
[47] V. M. Vokkarane, J. P. Jue, and S. Sitaraman, "Burst Segmentation: An Approach for Reducing Packet Loss in Optical Burst Switched Networks," in *Proc.* ICC 2002, New York, 2002.
[48] V. M. Vokkarane and J. P. Jue, "Prioritized routing and burst segmentation for QoS in optical burst-switched networks," in *Proc. Optical Fiber Communication Conf. (OFC) 2002*, Anaheim, CA, Mar. 2002.
[49] J. S. Turner, "Terabit burst switching," *J. High Speed Networks*, 1999.
[50] G. I. Papadimitriou, P. Tsimoulas, M. S. Obaidat, and A. S. Pomportsis, *Multiwavelength Optical LANs*. New York: Wiley, 2003.
[51] G. I. Papadimitriou and D. G. Maritsas, "Learning automata-based receiver conflict avoidance algorithms for WDM broadcast-and-select star networks," *IEEE/ACM Trans. Networking*, vol. 4, pp. 407–412, June 1996.
[52] G. I. Papadimitriou and A. S. Pomportsis, "Self-Adaptive TDMA protocols for WDM star networks: A learning-automata-based approach," *IEEE Photon. Technol. Lett.*, vol. 11, pp. 1322–1324, Oct. 1999.
[53] OMM Inc. [Online]. Available: http://www.ommInc.com/products/datasheets.html
[54] JDS Uniphase [Online]. Available: http://www.jdsu.com
[55] Lynx Photonic Networks [Online]. Available: http://www.lynx-networks.com/prod/packet8x8detail_new.asp
[56] NEL NTT Electronics Corp. [Online]. Available: http://www.nel-world.com/products/photonics/thermo_opt_88.html
[57] SpectraSwitch [Online]. Available: http://www.spectraswitch.com/products/product_line/wavewalker_2 × 2_datasheetnew.htm
[58] Agilent Technologies [Online]. Available: http://www.agilent.com/cm/photonicswitch/32 × 32_Product_Brief3_1_02.pdf
[59] Light Management Group, Inc. [Online]. Available: http://www.lmgr.net/data05.htm

Georgios I. Papadimitriou (M'89–SM'02) received the Diploma and Ph.D. degrees in computer engineering and informatics from the University of Patras, Patras, Greece, in 1989 and 1994, respectively.

From 1989 to 1994, he was a Teaching Assistant at the Department of Computer Engineering, University of Patras, and a Research Scientist at the Computer Technology Institute, Patras, Greece. From 1994 to 1996, he was a Postdoctorate Research Associate at the Computer Technology Institute. From 1997 to 2001, he was a Lecturer at the Department of Informatics, Aristotle University, Thessaloniki, Greece. Since 2001, he has been an Assistant Professor at the Department of Informatics, Aristotle University. He is a Member of the Editorial Board of the *International Journal of Communication Systems* and an Associate Editor of *Simulation: Transactions of the Society for Modeling and Simulation International*. He is coauthor of the books *Wireless Networks* (New York: Wiley, 2003) and *Multiwavelength Optical LANs* (New York: Wiley, 2003). He is the author of more than 80 refereed journal and conference papers. His research interests include design and analysis of optical and wireless networks and learning automata.

Chrisoula Papazoglou received the B.S. degree in computer science from the Department of Informatics of Aristotle University, Thessaloniki, Greece, in 2001. She is currently working toward the Ph.D. degree in optical communication networks at the same university.

Her research interests include wide area optical networks and optical switching.

Andreas S. Pomportsis (M'00) received the B.S. degree in physics and the M.S. degree in electronics and communications from the University of Thessaloniki, Thessaloniki, Greece, and a Diploma degree in electrical engineering from the Technical University of Thessaloniki, Thessaloniki, Greece. He received the Ph.D. degree in computer science from the University of Thessaloniki in 1987.

Currently, he is a Professor at the Department of Informatics, Aristotle University, Thessaloniki, Greece. His research interests include computer networks, computer architecture, parallel and distributed computer systems, and multimedia systems.