Architectures and Protocols that Enable New Applications on Optical Networks

Malathi Veeraraghavan, Ramesh Karri, and Tim Moors, Polytechnic University
Mark Karol, Avaya Inc.
Reinette Grobler, University of Pretoria

ABSTRACT

This article first discusses how advances in networking architectures and protocols can complement advances in optical communications research to increase the overall value of optical networks by enabling more applications. A review of existing optical networking solutions is then provided along with a classification of different types of optical networks. Finally, we show how single-hop and multihop wavelength-routed networks can be used efficiently for fast end-to-end file transfers when these networks are equipped with a hardware-implementable signaling protocol, a routing protocol, and a simple transport protocol.

INTRODUCTION

Advances in networking architectures and protocols are driven by both new inventions in communications technologies and new applications. The work presented in this article describes new optical networking architectures and protocols enabled by recent advances in multichannel wavelength-division multiplexed (WDM) communications. We will describe some of the important *optical communications components* such as WDM multiplexers/demultiplexers, programmable optical add/drop multiplexers (OADMs), optical crossconnects (OXCs), tunable and fixed transmitters, continuous-mode and burst-mode receivers (tunable or fixed), amplifiers, and WDM passive star couplers.

Besides these technological advances in optical communications components, optical fiber deployment is now finally reaching customer premises buildings (i.e., enterprise/multiple tenant buildings in metropolitan areas). Given that fiber was already deployed on long-haul lines (between cities/continents), end-to-end fiber connectivity is now available to many business users.

To exploit this increasing reach of fiber and the capabilities of new optical communications components, new networking architectures and protocols are needed to enable a larger set of communications applications on optical networks. A number of network architectures have already been proposed and some even implemented. As a review, we provide a *description/classification* of these networking solutions in a later section.

While these existing optical networking solutions make a good start, we believe that by adding a few key protocols, such as signaling protocols, routing protocols, and transport protocols, we can further improve the value of optical networks. Initial proposals for signaling and routing protocols for optical networks have been made, such as multiprotocol lambda switching by the Internet Engineering Task Force (IETF) [1], the Optical Domain Service Interconnect (ODSI) coalition [2], Optical Internetworking Forum (OIF) [3], and the International Telecommunications Union — Telecommunication Standardization Sector (ITU-T). These proposals allow optical networks to operate in a switched mode (i.e., for lightpaths to be set up and released on demand). However, the applications envisioned for optical networks operating in the switched mode are so far fairly limited. For example, the most common application is restoration following link failures. Optical networks are assumed to interconnect Internet Protocol (IP) routers, and when a failure occurs, an IP router requests a new lightpath to route around the failure. This application is targeted at providing reliability without requiring provisioned protection-switched paths (which waste bandwidth).

Figure 1 shows how new networking architectures and protocols can help increase the value of optical networks by enabling new applications. For example, the creation of OADMs and OXCs followed by the development of signaling and routing protocols enabled the efficient network restoration application. We recognize that with signaling protocols enabling the switched mode of operation, even more applications are possible. As a new application for optical networks, we propose using switched lightpaths on an end-to-end basis between hosts (instead of just

Reinette Grobler was funded by CATT, Center for Advanced Technology in Telecommunications, Polytechnic University, New York.

0163-6804/01/$10.00 © 2001 IEEE

between IP routers) for large file transfers. For large files, the overhead of lightpath setup is offset by savings in transmission delay relative to that in packet-switched networks, especially with congestion control schemes, such as Transmission Control Protocol (TCP) Slow Start. To support this application, we need a slightly different set of signaling and routing protocols as well as a new transport protocol. We describe our proposals for these new networking protocols and the associated application. This extension is depicted as a loop leading from the applications box into the networking architectures and protocols box of Fig. 1.

As a next step, we recognize that for this end-to-end file transfer application, fast reconfiguration capability of OADMs/OXCs would be most useful. It would improve circuit setup delay, and consequently reduce the crossover file size at which circuit switching becomes more efficient than packet switching. Currently, most emerging all-optical crossconnects use micro-electromechanical systems (MEMS) technology, in which reconfiguration of a lightpath takes in the order of milliseconds. New crossconnect technologies based on semiconductor optical amplifiers (SOAs) would reduce this to nanoseconds [4]. This is an example of the reverse loops leading into the optical communications components box of Fig. 1.

Further advances in optical communications components could enable more applications, if corresponding networking architectures/protocols are developed. For example, the creation of All-wave fibers opened up the band between 1350 and 1450 nm [5]. This allows for a significant increase in the number of wavelengths supported per fiber over previous technology. With such advances, if bandwidth becomes truly abundant and inexpensive, one could consider using switched bidirectional end-to-end circuits even for interactive sessions. Contrast this with the proposal discussed earlier of using switched unidirectional end-to-end circuits only for large file transfers, an application in which continuous traffic (no silence periods) is generated. If bandwidth becomes plentiful, and circuits can be set up and released dynamically between end hosts, interactive applications such as Web browsing and telnet could also be considered for end-to-end circuits. Traffic would be bursty, which means the circuits would be used inefficiently, but delays would be much lower than in bidirectional sessions set up across packet-switched networks.

Finally, advances in optical memory components could further improve the value of optical networks. For example, optical packet switches would become practical, which in turn would increase the number of applications supported on optical networks.

Thus, using this cyclical model of Fig. 1, we believe that both the deployment and value offered by optical networks can be well improved.

Optical Components

This section briefly describes the various optical communications components listed earlier. *Multiplexers* aggregate multiple wavelengths onto a single fiber, and *demultiplexers* perform the reverse function. These are typically static devices. *OADMs* are programmable in that they can be configured to add or drop different wavelengths. The whole multichannel signal does not need to be demultiplexed in an OADM, unlike in an *OXC*, where multiple fibers, each carrying multiple channels, are first terminated on demultiplexers before being crossconnected in a space-division switch fabric. These crossconnects are sometimes referred to as *wavelength routers* or *wavelength crossconnects*.[1] If wavelength conversion capability is present, the crossconnects are referred to as *wavelength interexchange crossconnects* (WIXCs). Otherwise, they are referred to as *wavelength selective crossconnects* (WSXCs). *Tunable transmitters* have either lasers whose output wavelength can be tuned as needed, or an array of lasers with different wavelengths that can be selectively enabled, while *fixed transmitters* have lasers whose wavelengths are set during manufacturing. *Burst-mode receivers* can synchronize to a transmitter's signal very rapidly, allowing their use with transmitters that send bursts of data (alternating with silences), unlike *continuous-mode receivers* that have slow synchronization times and hence require the transmitter to send a continuous signal. Both types of receivers can be tunable or fixed in the wavelengths they can receive. Advances in *amplifier* technology have increased the distances between signal regenerators. *WDM passive star couplers* are simply broadcast devices that mix all the input signals and broadcast these on to all outgoing fibers. Finally, *optical packet switches* are nodes that have optical buffering capability and perform the packet header processing functions required of packet switches. This set of optical components serves as a sample set; other types of components have been proposed and implemented in the past.

We classify the components described above into two types: switching components, which are programmable and hence enable networking, and nonswitching components, which are primarily used on optical links. Switching components include tunable transmitters/receivers, OADMs/OXCs (including WSXCs and WIXCs), which are optical circuit switches, and optical packet switches. The remaining components, WDM multiplexers/demultiplexers, fixed transmitters/receivers, amplifiers, and WDM passive star couplers, are all nonswitching components.

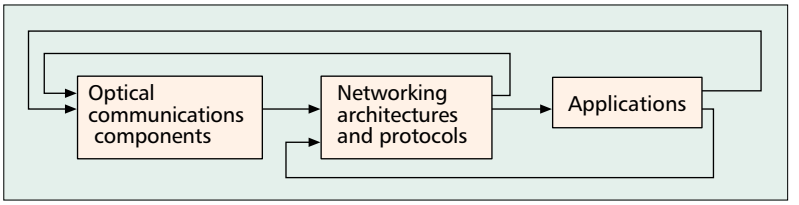

Figure 1. *Toward advancing the value of optical networks.*

[1] *These wavelength crossconnects can be "all-optical" if there is no conversion to the electronic domain, or "electronic" if there is conversion to the electronic domain. We refer to the former generically as "WDM crossconnects," but add the adjective "electronic" for the latter case.*

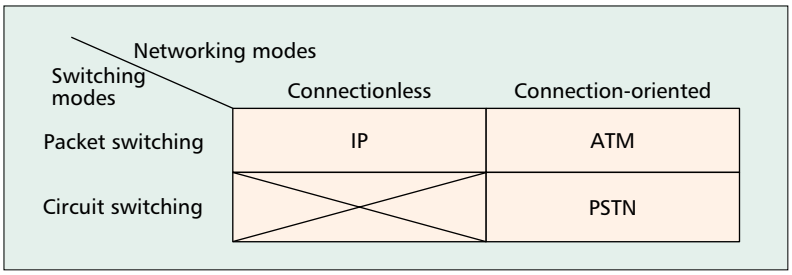

Figure 2. *Classification of some networking techniques.*

Finally, if wavelength converters are added as either separate components or WIXCs, a programmable converter is a switching component, while a fixed-wavelength converter is a nonswitching component.

Classes of Optical Networks

While commercial interest in WDM is relatively new, a significant amount of work has been done on optical networking in both universities and research laboratories over the last few decades. Recently published books [6–8] provide excellent reviews of this work. In this section we briefly describe different classes of optical networks both proposed in prior research papers and currently being deployed.

Before considering classes of optical networks, we consider classes of generic networks, whether electronic or optical. Figure 2 classifies networks based on their networking modes and switching modes. The networking mode is primarily either *connection-oriented* (CO) or *connectionless* (CL). CO networks are those in which connection setup is performed prior to information transfer. In contrast, in CL networks no explicit connection setup actions are performed prior to transmitting data; instead, data packets are routed to their destinations based on information in their headers. The switching mode of a network indicates whether the network nodes are *circuit switches* or *packet switches*. Circuit switches are position-based, in that bits arriving in a certain position are switched to a different output position, with the position determined by a combination of one or more of three dimensions: space (interface number), time, and wavelength. Packet switches are label-based, in that they use information in packet headers (labels) to decide how to switch a packet. Note that our definition of circuit switching is different from a common definition, which is that a circuit-switched network is one in which a circuit is set up prior to user data exchange. While this property is a characteristic of circuit-switched networks, it is not the defining property, because with the invention of packet-switched CO networking such as X.25 and asynchronous transfer mode (ATM), even in packet-switched networks a connection can be set up prior to data exchange. Whether a connection is set up prior to data exchange or not is a property of whether the network is CO or CL, not whether the network is packet- or circuit-switched. Figure 2 provides examples of the three networking techniques (CL, circuit-switched CO, and packet-switched CO). An IP network is an example of a CL packet-switched network. While we recognize that the addition of Resource Reservation Protocol (RSVP) and/or multiprotocol label switching (MPLS) adds a CO mode of operation to IP networks, for the purposes of this article we will use ATM as an example of CO packet-switched networks and IP as an example of CL packet-switched networks. The fourth networking technique (CL circuit switching) is not possible because circuit switches operate based on the position of arriving bits, which means that they must be programmed in a connection setup procedure.

Returning now to the classification of *optical networks*, we first determine which of the three networking techniques shown in Fig. 2 have been implemented using the optical communications components described earlier. The goal of carrying out this exercise is to determine if any known networking technique is as yet unused in the optical domain. Different types of optical networks, such as broadcast-and-select networks, wavelength-routed networks, optical link networks, single-hop networks, multihop networks, and photonic networks, have been described in [6, 7, 9]. Many of these terms have multiple (sometimes inconsistent) definitions. Developing a taxonomy for classes of optical networks is a challenging proposition. Nevertheless, it is important given the growth of the optical community in recent years.

We define four broad classes of optical net-

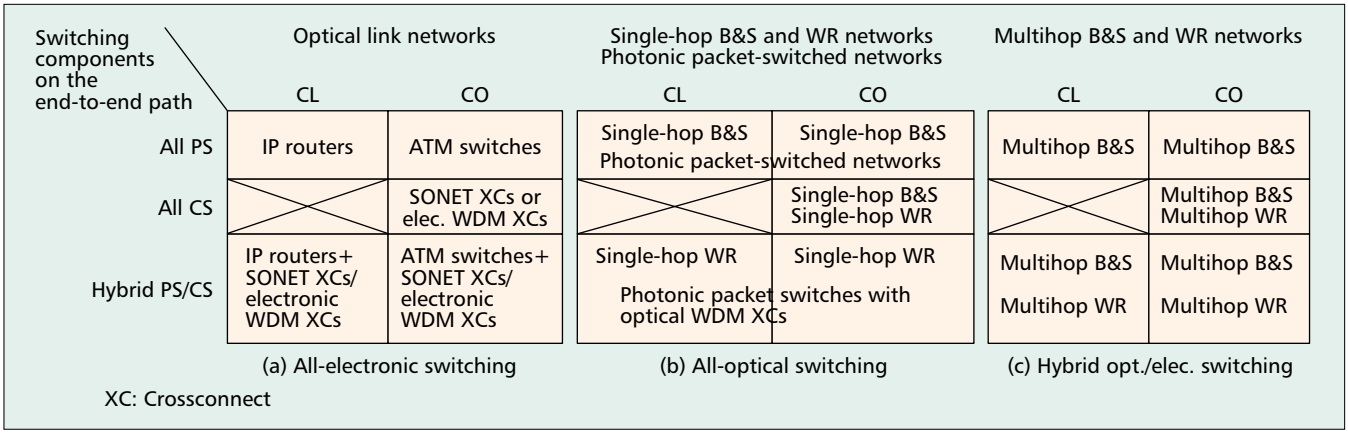

Figure 3. *Classification of optical networks.*

	Classes of optical networks			
Types of optical communications components	**Optical link networks**	**Broadcast-and-select networks**	**Wavelength-routed networks**	**Photonic packet-switched networks**
Nonswitching optical components	√	√	√	√
Tunable transmitters and/or tunable receivers	X	√	May or may not be present	May or may not be present
Optical circuit switches (OADMs and OXCs)	X	X	√	May or may not be present
Optical packet switches	X	X	X	√

Table 1. *Usage of optical communications components in optical networks.*

works based on the types of optical communications components used (Table 1). We assume that all classes use optical links and have nonswitching optical components (i.e., optical link components, e.g., amplifiers, fixed transmitters/receivers, WDM multiplexers/demultiplexers, and WDM passive star couplers).

Optical link networks consist of all-electronic switches interconnected by optical links. The optical links can be single-channel or multichannel point-to-point links, or shared-medium broadcast links. Point-to-point multichannel links are created by placing WDM multiplexers/demultiplexers at the ends of the fiber, while shared-medium broadcast links are created through the use of WDM passive star couplers. These components are not programmable, and hence no reconfiguration is possible. Figure 3a, an extended version of Fig. 2, shows different subclasses of optical link networks based on the type of electronic switches used on an end-to-end path. They could be all packet switches (of CL or CO type, or a combination, e.g., MPLS, not shown in Fig. 3a), all circuit switches, or a hybrid of packet and circuit switches.

The next class of optical networks shown in Table 1 are *broadcast-and-select* (B&S) networks. The only optical switching components used in these networks are tunable transmitters and/or receivers. As the name suggests, data is broadcast on all the links (e.g., all outgoing links of a WDM passive star coupler), and receivers are programmed to select the channels they should receive. B&S networks are classified as either single-hop or multihop. The terms *single-hop* and *multihop* indicate whether user data only traverses optical switching components on the end-to-end path (single-hop) or whether it traverses a combination of optical and electronic switching components (multihop). This is important because all-optical networks have the major advantage of bit rate transparency.

In single-hop B&S networks, the transmitters/receivers can be tuned on a packet-by-packet or call-by-call basis. Thus, all three networking techniques, packet-switched (PS) CL, PS CO, and circuit-switched (CS), are theoretically possible in single-hop B&S networks, as shown in Fig. 3b. Hybrid CS/PS networks are not possible in this class because the only switching components used are tunable transmitters/receivers, and hence both ends of the end-to-end path need to be operated in the same mode, CS or PS.

In *multihop B&S networks*, data is broadcast on all links, but electronic switches (effectively) provide wavelength conversion on the path from the source to the destination because not all nodes receive all wavelengths. Given that these networks are B&S, the only optical switching components present are tunable transmitters/receivers (Table 1). Since the components can be tuned on a packet-by-packet or call-by-call basis and the electronic switches can be circuit or packet switches, multihop B&S networks can be operated in all categories of Fig. 3c (except the all-CS/CL category). Reference [6] classifies a network in which the transmitters and receivers are fixed and the end nodes are connected through a WDM passive star coupler as a multihop B&S network. We classify this network as an optical link network because in this configuration there are no optical switching components (note that we classified WDM passive star couplers as optical link components rather than as optical switching components in an earlier section). While the generic term multihop may allow for networks consisting of electronic switches interconnected by optical links, we reserve the term multihop B&S for hybrid networks consisting of both optical and electronic switching components (hence, we classified these networks in the hybrid category of Fig. 3c).

The third class shown in Table 1, *wavelength-routed* (WR) *networks*, are defined to necessarily include optical circuit switches (OADMs/OXCs), and optionally tunable transmitters and/or receivers. WR networks can also be single-hop or multihop. *Single-hop WR networks* use only optical switching components and are hence listed in Fig. 3b. If the transmitters/receivers are fixed, these networks are all-CS. With tunable transmitters/receivers [10], based on whether the transmitters/receivers are tuned on a per-packet or per-call basis, these networks are in the hybrid PS/CS (last row of Fig. 3b) or all-CS category (second row of Fig. 3b), respectively. *Multihop WR networks* consist of optical circuit switches, electronic switches, and optionally tunable transmitters/receivers. Since they necessarily include optical circuit switches, they are not listed in the all-PS row of Fig. 3c. The electronic switches on the multihop paths could be circuit-switched, or CO or CL packet-switched; hence, multihop WR networks are listed in the corresponding cate-

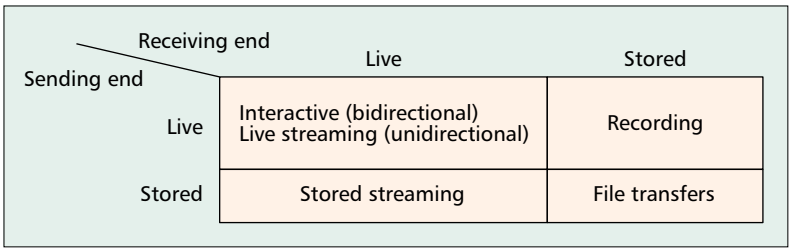

Figure 4. *Types of data transfers.*

gories in Fig. 3c.

The last class of optical networks shown in Table 1 are *photonic packet-switched networks*. These networks are defined to necessarily have optical packet switches, and optionally optical circuit switches and tunable transmitters and receivers. These networks are listed in Fig. 3b (all-optical switching) in the last hybrid row and in the all-PS row because optical circuit switching may or may not be present on the end-to-end path. Optical circuit switching could be in the form of OADMs/OXCs or tunable transmitters/receivers tuned on a call-by-call basis. While it is possible to combine all-optical packet switches with electronic circuit switches, we think that these configurations are unlikely because when all-optical packet switches become prevalent, all-optical circuit switches are already likely to have superseded electronic circuit switches. Hence, photonic packet-switched networks are not shown in Fig. 3c.

Finally, a new class of optical networks, not listed in Table 1, is based on *optical burst switching* [11, 12]. Burst switching combines the concepts of circuit and packet switching. While a round-trip call setup delay is incurred in circuit switching, burst switching avoids this cost by sending the burst at some time interval following the burst control packet, optimistically hoping that resources will be allocated for the burst before its arrival at a switch. The cost, however, is that given there is some probability of the burst arriving before resources are allocated, buffers are needed in burst switches.

Of all these classes of optical networks, clearly optical link networks are feasible today and are already in use. Of the remaining three classes of optical networks described above (B&S, WR, and photonic packet-switched networks), current commercial attention is directed at WR networks. The tunable transmitters and receivers needed for B&S networks, and optical memory needed for optical packet switches, remain quite expensive, making both B&S and photonic packet-switched networks less attractive for near-term commercial deployments.

Of the various subclasses of WR networks shown in Fig. 3b and c, the most common are multihop WR networks with electronic packet switches (i.e., the last row of Fig. 3c); for example, a network of IP routers interconnected by optical circuit switches, such as OADMs/OXCs. As described earlier, limited applications are possible for the WR subnetwork of this integrated IP/WDM network. In the next section, we describe our proposal for extending the value of WR networks.

A Proposal for New Architectures, Protocols, and Applications for WR Networks

In this section we propose using WR networks in two other subclasses: the all-CS case of the all-optical configuration (Fig. 3b) or the all-CS case of the hybrid electronic/optical switching configuration (Fig. 3c). We use synchronous optical network (SONET) switches as representative of electronic circuit switches, and use the term *WDM switches* to represent all-optical OADMs and OXCs. Thus, all-CS single-hop WR networks consist only of WDM switches, while the all-CS hybrid electronic/optical configuration uses SONET and WDM switches. Note in Fig. 3 that rows corresponding to the all-CS case imply that all the switching components on the end-to-end path use circuit switching.

To understand what types of data transfers are suitable for end-to-end circuits, we considered different types of data transfers as shown in Fig. 4. We emphasize that Fig. 4 shows a classification of data transfers and not applications. A communication application session can have different types of data transfers. Each data transfer has two ends (assuming two-party sessions, for simplicity), sending and receiving. Each of these ends can be live or stored. Live does not imply the presence of a human user. A computer could consume received data live as part of a distributed application. Examples of live-to-live sessions are interactive telephony and videoconferencing, which are bidirectional, and live streaming, which is unidirectional. An example of live streaming is a live radio broadcast. An example of the recording type is the recording of a live TV broadcast. Stored streaming examples include a user listening to a stored audio file being streamed over a network or a video-on-demand data stream. All three types in which one end is live typically involve bursty data. Only the last type, stored-to-stored, involves continuous traffic. Hence, we conclude that this last type of data transfer is best suited to end-to-end circuits (WDM lightpaths or SONET circuits).

Given that such data transfers are often part of other interactive sessions (e.g, a file download within a Web browsing session), we propose a hybrid network architecture in which short requests, such as uniform resource locators (URLs), are carried over a CL packet-switched network such as an IP network, while long file transfers, if needed, are transmitted over a direct end-to-end circuit. This is illustrated in Fig. 5. We assume a gradual introduction of SONET and/or WDM switches into enterprises, and SONET or WDM network interface cards (NICs) at end hosts. The network example shown in Fig. 5 could be a single-hop WR network with only optical switches and/or a multihop WR network with optical (WDM) and electronic SONET switches. Many large enterprises are already deploying next-generation SONET equipment (multiplexers/crossconnects) within customer premises buildings to aggregate

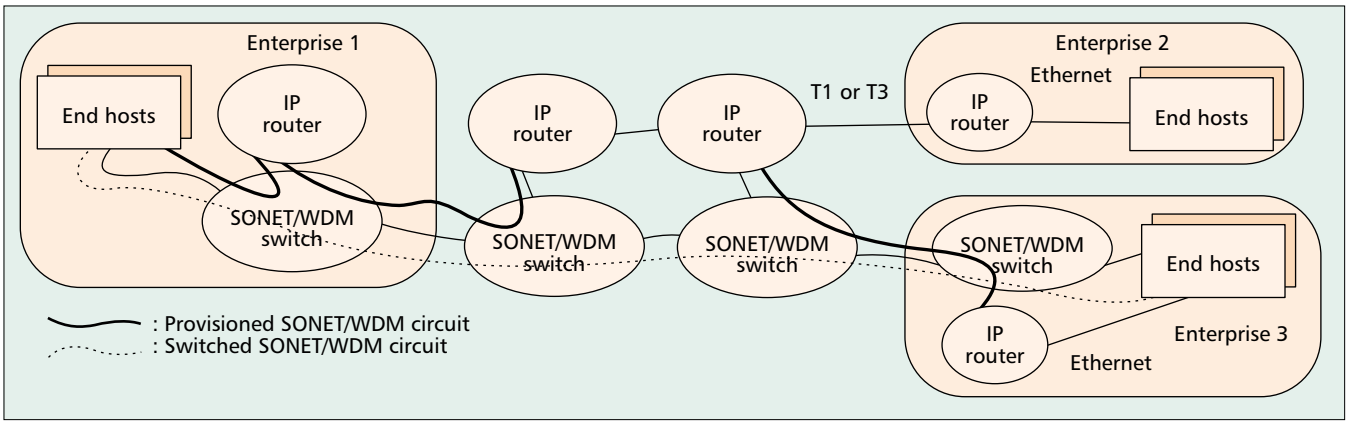

Figure 5. *The proposed hybrid architecture for file transfers.*

voice and data traffic onto SONET optical circuits. Also, building local exchange carriers (BLECs), a newly emerging class of service providers, have added optical fiber cabling within large enterprise and multiple tenant buildings, making it feasible to take fiber all the way to desktops. IP routers in the enterprises are connected to IP routers in the wide area network (WAN) via provisioned SONET/WDM circuits as shown in Fig. 5. End hosts can connect to enterprise routers through SONET/WDM provisioned circuits as in enterprise 1 of Fig. 5, or via Ethernet as in enterprise 3. Additionally, an end host can request an end-to-end circuit on demand for large file transfers (as shown by the dotted line in Fig. 5) if the far end also has a SONET/WDM NIC. Thus, a host would use the IP network to send a file request (e.g., a URL), but the file, if deemed large, would be sent on a direct end-to-end SONET/WDM circuit. For large files, circuit setup overhead could be smaller than overheads associated with packet headers, acknowledgments, and congestion control mechanisms in packet-switched networks.

To reduce circuit setup delay, we designed a simple signaling protocol suitable for hardware implementation and implemented it in VHDL.[2] Our first prototype is quite promising, showing significant savings in call setup delays. This is described below. We will also describe a transport protocol called *Zing* that uses this hybrid network architecture for reliable delivery of files on end-to-end circuits. We will describe a routing protocol for single-hop and multihop WR networks using circuit switches. Finally,we describe an interesting problem that arises when admitting calls in an end-to-end circuit-switched network, and propose a solution.

Hardware Implementation of a Signaling Protocol

Circuit-switched networks are, by definition, connection-oriented (Fig. 2). CO networks require a signaling protocol to set up and release connections. Typically, signaling protocols are implemented in software, which both limits the call handling capacities of switches and leads to high call setup delays. Call setup delays are high relative to call holding times if calls are set up and released for the purposes of transferring files (unidirectional) on potentially very high-speed circuits. Thus, to support the application described above, that is, using a CL packet-switched network for short requests within a Web browsing or ftp session and using direct high-speed circuits for file transfers (whenever the files are deemed large), the switches need to have high call handling capacities and very low call setup delays. Our solution to this problem is to implement the signaling protocol in hardware.

To date, hardware implementations of signaling protocols have not been considered because of two reasons:

- Complexity of signaling protocols
- Flexibility to support evolving protocols

On the first count, we contend that signaling protocols for circuit-switched networks can be simple enough for hardware implementation. Unlike for PS CO networks, where signaling protocols support many traffic descriptor parameters to model variable bit rate traffic, and many QoS parameters, such as packet loss, delay, and delay variation, signaling protocols for circuit-switched networks need only deal with one metric (i.e., bandwidth). The usual complexity associated with state management can also be handled by viewing state table manipulation as simple read/writes of a state table. The second count, hardware inflexibility, is somewhat invalidated by the concept of reconfigurable hardware. Examples of reconfigurable hardware include field programmable gate arrays (FPGAs). As the name suggests, as signaling protocols are upgraded, hardware implementations can be downloaded to the FPGAs in the field.

To validate this line of thinking, we designed a simple signaling protocol for circuit-switched networks and implemented it in VHDL. There are only four signaling messages: *Setup*, *Setup-Success*, *Release*, and *Release-confirm*. Processing a signaling message primarily entails manipulating data tables at each switch along the end-to-end path. For example, when a *Setup* message arrives, the signaling protocol engine consults a *routing table* to determine the next hop switch, and an *available capacity* table to determine if the call can be admitted. If the call is admitted, it writes a *switch*

[2] *VHDL stands for VHSIC Hardware Description Language, where VHSIC stands for very high-speed integrated circuit.*

The prototype validated the feasibility of hardware implementation of signaling protocols, and its potential for achieving very high call handling capacities and low call setup delays.

configuration table to indicate how user data bits should be switched after call setup, and a *state table* indicating the state of the call.

We compiled, simulated, and synthesized a prototype VHDL model of the signaling protocol engine using Altera Maxplus II 8.3 FPGA design tools. The design fits into an Altera FLEX 10K100GC503-3 FPGA device with 60 percent resource utilization (about 30,000 gates).

From the timing simulation, we determined that receiving/transmitting a *Setup* message consumes 9 clock cycles, while processing the *Setup* message consumes 25 clock cycles (this includes a match time of 5 clock cycles for the routing table, a match time of 3 clock cycles for the available capacity table based on a Motorola MCM69C432 CAM device with a worst case match time of 8 clock cycles). Overall, 45 clock cycles are necessary to receive and process a *Setup* message. Processing *Setup-success*, *Release*, and *Release-confirm* messages consumes about 15 clock cycles each since these messages are much smaller and require simple processing. Assuming a 25 MHz clock, this translates into 1.8 μs for *Setup* message processing and about 0.6 μs for *Setup-success*, *Release*, and *Release-confirm* message processing. Compare this with the 1–2 ms it takes to process signaling messages in software [13]. FPGAs with 100 MHz clocks are already on the market, and using them will reduce the protocol processing time even further. Such low call setup delays also imply high call handling capacities on the order of 100,000 calls/s.

The prototype validated the feasibility of hardware implementation of signaling protocols, and its potential for achieving very high call handling capacities and low call setup delays. This technological advance indeed enhances the set of applications that can be supported efficiently on optical networks.

TRANSPORT PROTOCOL

Once an end-to-end circuit is set up, a transport protocol is needed to reliably carry user data bits on the circuit. Since the end-to-end circuit can be considered a link at the physical layer, the network layer is nonexistent, and the transport layer and data link layer protocols are effectively merged. With end-to-end circuits, data loss probability is much lower than in packet-switched networks, where data can be dropped at intermediate switches due to buffer overflows. Also, data is not missequenced. This means a simpler error control mechanism than that of TCP is sufficient. Also, without the threat of network switch congestion, data can be sent at a constant rate without congestion control functionality, such as TCP's Slow Start. Finally, without the variations in data delivery rates commonly experienced in CL packet-switched networks, flow control does not have to be the fine-grained window-based control used in TCP. A simpler rate-based scheme is possible on circuits.

Given the above, we defined a new merged transport/data link layer protocol called Zing [14]. It mainly performs two functions: *error control* and *flow control*. Typically these functions are common to both transport protocols (on an end-to-end basis) and data link layer protocols (on a link-by-link basis). Other functions unique to one layer or the other (e.g., framing typically performed in data link layer protocols) are also needed.

For *error control*, we need to define a specific unit of data over which error detection and correction functions can be performed, even though on an end-to-end circuit, data bits can be streamed continuously. In Zing, this transmission unit is called a *chunk*. Chunks are of fixed size, except for the last. Each chunk consists of data payload bits encapsulated with a sequence number and a checksum. The Zing receiver verifies correct reception of each chunk (using the checksum), and delivers the chunk to the storage device (e.g., disk). If the receiver detects a chunk error, it sends a negative acknowledgment to the source to request retransmission. By allocating sequence numbers to chunks rather than bytes, using a large sequence number field (32 bits) and large chunk sizes (such as 64 kbytes), Zing avoids having to recycle sequence numbers within the duration of a file transfer. Consequently, Zing can avoid retransmission buffers at the source and resequencing buffers at the destination, by asking the application to:

- Retrieve information from disk when needed for retransmission
- Store information to disk as it arrives, possibly with holes

This approach is in contrast to stream-oriented transport protocols, such as TCP, that provide a serial interface to the application, and consequently hasten retransmissions to release storage in the limited-size retransmission and resequencing buffers. With Zing, the destination can issue a negative acknowledgment when it detects that information needs to be retransmitted (a simple process given that circuits preserve sequence), and the source can retransmit the information much later. With this approach, Zing exploits the fact that files must be complete before any part can be used, and files are stored (e.g., on disk) at the endpoints.

For *flow control* functionality, Zing uses a rate-based scheme, that is, the rate at which a Zing receiver can receive data is taken into account when setting up the end-to-end circuit. The other two commonly used mechanisms for flow control, window-based flow control (as in TCP) and on-off flow control (as in high-level data link control), are not used for various reasons. Window-based flow control is more complex, requiring the end nodes to manage window sizes. On-off flow control is unsuitable for high-speed circuits in that during OFF periods, rather large amounts of bandwidth will go unused.

While Zing is designed for file transfers over a unidirectional circuit-switched lightpath, it uses the packet-switched network that operates bidirectionally in parallel with the circuit. Zing uses the packet-switched network for sending negative acknowledgments. More interestingly, Zing uses the packet-switched network for retransmissions. Packet-switched retransmissions are desirable for two reasons. First, they allow Zing sources to efficiently close the circuit immediately after transmitting the last bit of the file for the first time. If Zing sources were to retransmit over the circuit, they would need to hold the cir-

cuit open and idle until receiving acknowledgments for all bits in the transfer, which could not occur until at least a round-trip time after the last bit of the file was transmitted (and might take much longer due to the tolerances required on timers that control timeouts). Second, packet-switched retransmissions constrain the circuit holding times to known values, a feature we need for our call scheduling algorithm, which is described later. The number of chunk retransmissions needed depends on the error rate experienced during the transfer, which means retransmission on the circuit could make the circuit holding time unknown.

Reliability is of concern not just for the initial transmissions, but also for the acknowledgments and retransmissions sent over the packet-switched network. Rather than have Zing reinvent a reliable transfer protocol for the packet-switched network, we propose using TCP. This simplifies Zing, allowing initial transmissions, which constitute the bulk of the transfer, to zip across the high-speed circuit/lightpath, and relatively rare acknowledgments and retransmissions to pass slowly but reliably across the packet-switched network.

Zing also addresses the denial of service attacks made possible by the choice of negative acknowledgments, and the impact of out-of-order access to storage on encryption modes.

Routing Protocol

Current circuit-switched WR networks, whether single-hop or multihop, do not have a routing protocol. Consider, for example, SONET networks. These are typically deployed in rings. Example ring rates are OC-12, OC-48, OC-192, and so on, and individual OC-1s, OC-3s, and so forth are dropped/added at SONET add/drop multiplexers (ADMs). Similarly, WR networks with OADMs are also emerging as 4-channel, 16-channel, 32-channel, and so on WDM rings. The advantages of such network configurations are:

- ADMs are cheaper than crossconnects given their ability to drop individual channels without demultiplexing the complete signal.
- Leased fiber costs are lower in a ring configuration since fiber drops (which are typically charged on a per mile basis) are between consecutive ADM locations rather than between a customer premises building and some centralized service provider hub.

However, the disadvantage of these ring configurations is that if any one customer wants to upgrade his/her service to a higher rate than the ring rate, a complete upgrade of the ring is required.

To overcome this drawback, if one proposes using crossconnects instead of ADMs so that different users can have interfaces at different rates, the cost of demultiplexing becomes an issue. For example, a crossconnect operating at the OC1 rate would need to demultiplex all incoming interface signals to OC1 signals before crossconnecting, which increases line card costs. One way to alleviate this problem is to support crossconnects operating at various rates. To allow SONET crossconnects at OC-1, OC-3, OC-12, OC-48, and OC-192 rates to coexist with each other and with OXCs operating at wavelength granularity, the routing and signaling protocols need features to support such heterogeneous networks.

Thus, our first requirement for the routing protocol (and associated signaling protocol) is to support *heterogeneous networks*. The routing protocol would allow the electronic and/or optical circuit switches to not only exchange topology and loading information, but also their crossconnect rates. Based on this information, a route precomputation module would fill a routing table with next hop node information for circuits at different rates. If shortest paths happen to be through a set of switches that have higher crossconnect rates than the desired circuit rate, the routing table would provide sufficient information to allow the signaling protocol processor to first set up a higher-rate circuit between two low-rate switches, and then continue the set up of the lower rate (desired) circuit. Details of the mechanism are currently under development.

Another aspect that makes the routing protocol challenging is the need to allow subnetworks consisting solely of WSXCs (i.e., nodes *without wavelength conversion capability*). The routing protocol messages then have to carry lists of available wavelengths at downstream nodes to upstream nodes to allow the latter to select a wavelength for a new call that has a high probability of success.

More common questions that need to be answered in designing a routing protocol are whether to use distance-vector or link-state schemes, hop-by-hop or source routing, flat or hierarchical routing, and so on. Currently, we have selected a link-state scheme, hop-by-hop routing, and a two-level hierarchical network. *Link-state mechanisms* are better when multiple metrics are considered; in this case, link weights and available bandwidth. The latter is important in routing protocols for circuit-switched networks unlike in CL packet-switched networks because, in CS networks, a call will get rejected if the route taken does not have sufficient bandwidth, while in CL PS networks all calls will be admitted even if bandwidth is constrained. In the latter case, packets may be dropped on congested paths but recovered through error control mechanisms. We chose *hop-by-hop routing* over source routing even though the latter decreases the probability of loops that can arise when available bandwidth is considered. This decision was made to keep the signaling protocol simple for hardware implementation. A *two-level hierarchical scheme* was chosen over a flat scheme for scalability reasons.

Thus, given the need to support heterogeneous networks, the wavelength continuity constraint, and keeping the signaling protocol simple (to enable hardware implementation), no existing routing protocol could be readily reused for these all-CS WR networks.

Given the need to support heterogeneous networks, the wavelength continuity constraint, and keeping the signaling protocol simple (to enable hardware implementation), no existing routing protocol could be readily reused for these all-CS WR networks.

The Call Scheduling Algorithm

In this section we describe an interesting problem that arises when end-to-end circuits are used for file transfers. In today's Internet, since file transfers are viewed as not having stringent delay requirements, they are often carried out

A new connection admission control mechanism is required to achieve our goal of low call blocking probabilities and high utilization. It turns out that such a scheme can indeed be designed if calls have known holding times.

in a mode that achieves high network utilization while sacrificing delay. In a TCP/IP network, a file transfer request is as such never rejected; however, packets of the file transfer often experience large delays. This is in contrast to CO networks, which usually operate in a call blocking mode. When a request for a connection arrives in a telephony, ATM, or MPLS network, the call is blocked if the requested resources are not available. To guarantee low call blocking probabilities in these networks, overprovisioning of resources is required, which results in low utilizations. Our question then was to determine whether there was a way to keep call blocking to a minimum (zero if possible), but incur a call start delay in return for improved utilization. Hence, we considered call queuing schemes.

Allowing switches to queue calls is a way to improve network utilization in circuit-switched networks, i.e., by using buffers to hold call setup requests, there is a greater assurance of keeping network resources in use. However, a problem arises when calls are queued in sequence at switches on the end-to-end path waiting for network resources. The problem is that while a call setup request is queued at a downstream switch for resources, upstream resources that were assigned to the call are idle. Sequential queuing will not only result in low utilization; interestingly, it also increases the call blocking probability over that in a network that implements pure call blocking.

A new connection admission control mechanism is required to achieve our goal of low call blocking probabilities and high utilization. It turns out that such a scheme can indeed be designed if calls have known holding times. Luckily, this is true for file transfers on circuits, where the call holding times are known. Call holding time for a file transfer consists of the file transmission time (file size plus overhead divided by circuit rate) and propagation delay of the circuit medium.

With knowledge of call holding times, a switch can maintain a time-varying available capacity function for each outgoing interface. This function reflects the scheduled start times of all the connections on that interface. When a new call arrives, each switch provides a delayed start time. This allows for a call to be admitted at multiple switches of an end-to-end path without impacting utilization. We designed two new connection admission control algorithms, simulated them, and showed that high utilizations can indeed be achieved while keeping call blocking probabilities low at the cost of call start time delays [15].

SUMMARY

The two main contributions of this article are as follows. First, we provide a classification for the broad set of optical networks proposed in research literature and commercial deployments. Next, we illustrate the importance of creating key networking protocols to augment the basic capabilities provided by optical communications components. Specifically, given the commercial interest in wavelength-routed networks, we demonstrate that with appropriate new networking protocols, optical networks can be used for far more applications than they currently are. We propose a hybrid architecture consisting of a connectionless (packet-switched) network, such as an IP network, and a parallel all-circuit-switched WR network (end-to-end). We demonstrate how common Internet applications, such as Web browsing, could take advantage of a high-speed circuit-switched network for file transfers on end-to-end circuits. By implementing a signaling protocol in hardware, we propose to support call setups and releases for individual file transfers at high call handling rates and with low call setup delays. A new transport protocol, called Zing, is proposed for the end-to-end circuit in this hybrid network. Finally, we address important call scheduling and routing protocol issues to achieve high network utilization and flexible architectures.

REFERENCES

[1] D. Awduche *et al.*, "Multi-Protocol Lambda Switching: Combining MPLS Traffic Engineering Control with Optical Crossconnects," draft-awduche-mpls-te-optical-02.txt, July 2000.

[2] A. Copley, "Optical Domain Service Interconnect (ODSI): Defining Mechanisms for Enabling On-Demand High-Speed from the Optical Domain," *IEEE Commun. Mag.*, vol. 38, no. 10, Oct. 2000, pp. 168–74.

[3] D. Pendarikis, B. Rajagopalan, and D. Saha, "Routing Information Exchange over the UNI and NNI," *Proc. OIF 2000*, Apr. 2000.

[4] M. Bachmann *et al.*, Polarization-insensitive Clamped-gain SOA with Integrated Spot-size Converter and DBR Gratings for WDM Applications at 1.55 mm Wavelength," *IEEE Elect. Lett.*, vol. 32, no. 22, Oct. 24, 1996, pp. 2076–78.

[5] J. J. Refi, "Optical Fibers for Optical Networking," *Bell Labs Tech. J.*, vol. 4, no. 1, Jan.–Mar. 1999, pp. 246–61.

[6] B. Mukherjee, *Optical Communication Networks*, McGraw Hill,1997.

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

[8] T. Stern, *Multiwavelength Optical Networks: A Layered Approach*, Addison-Wesley, 1999.

[9] J. Bannister, M. Gerla, and M. Kovacevic, "Routing in Optical Networks", Ch. 7, *Routing in Communication Networks*, M. Steenstrup, Ed., Prentice Hall, 1995.

[10] M. W. Janoska and T. Todd, "A Single-Hop Wavelength-Routed LAN/MAN Architecture," *Proc. IEEE Infocom '96*, San Francisco, CA, Mar. 24–28, 1996, pp. 402–9.

[11] C. Qiao and M. Yoo, "Optical Burst Switching (OBS) – A New Paradigm For An Optical Internet," *J. High Speed Networks*, vol. 8, no. 1, 999, pp. 69–84.

[12] J. Turner, "Terabit Burst Switching," *J. High Speed Networks*, vol. 8, no. 1, 1999, pp. 3–16.

[13] P. Pan and H. Schulzrinne, "YESSIR: A Simple Reservation Mechanism for the Internet," IBM res. rep. RC 20967, Sept. 2, 1997.

[14] T. Moors and M. Veeraraghavan: "Specification of (and Reasoning Behind) Zing: A Transport Protocol for File Transfers Over Circuits," CATT tech. rep., Polytechnic Univ.; http://uluru.poly.edu/~tmoors/zing

[15] R. Grobler, M. Veeraraghavan, and D. M. Rouse, "Scheduling Calls with Known Holding Times," CATT tech. rep., Polytechnic Univ.; http://kunene.poly.edu/~mv/ps-files/call-sched.ps.

BIOGRAPHIES

MALATHI VEERARAGHAVAN [SM] (mv@poly.edu) is currently an associate professor in the Department of Electrical Engineering at Polytechnic University. She received her B.Tech. degree in electrical engineering from Indian Institute of Technology, Madras, in 1984, and M.S. and Ph.D. degrees in electrical engineering from Duke University in 1985 and 1988, respectively. She worked for 10 years in Bell Laboratories conducting research on various networking protocols and control algorithms. She holds 16 patents, and has received four Best Paper awards. She served as an associate editor of *IEEE Transactions on Reliability* from 1992 to

1994. She is currently IEEE Communications Society E-News Editor, and an Area Editor for *IEEE Communication Surveys*.

RAMESH KARRI (ramesh@india.poly.edu) received a Ph.D. in computer science from the University of California at San Diego in 1993. During 1988–1989 he worked as a research engineer at CMC Ltd. in India. During 1993–1998 he was an assistant professor in the Department of Electrical and Computer Engineering at the University of Massachusetts at Amherst. During 1997–1998 he was member of technical staff at Bell Laboratories, Lucent Technologies. Since August 1998 he is an associate professor of electrical engineering at Polytechnic University, Brooklyn, New York. His research interests include hardware implementation of optical networking protocols, CAD for fault tolerance, reliability, and manufacturability, reconfigurable computing, and system-level power management. He has served as a guest editor for special issues of *IEEE Transactions on CAD*, *Reliability*, and *IEEE Design and Test Magazine*. He was the recipient of an NSF CAREER award.

TIM MOORS [M] (moors@ieee.org) is a research scientist at the Center for Advanced Technology in Telecommunications at Polytechnic University. He researches transport protocols for wireless and optical networks (e.g., Zing), wireless LAN MAC protocols that support bursty voice streams, communication system modularity, and fundamental principles of networking. Previously, he was with the Communications Division of the Australian Defence Science and Technology Organisation. He received his Ph.D. and B.Eng. (Hons.) degrees from universities in Western Australia (Curtin and UWA).

MARK J. KAROL [F] (mk@avaya.com) received a B.S. degree in mathematics and a B.S.E.E. degree in 1981 from Case Western Reserve University, and M.S.E., M.A., and Ph.D. degrees in electrical engineering from Princeton University in 1982, 1984, and 1986, respectively. From 1985 to 2000 he was a member of the Research Communications Sciences Division at Bell Laboratories, Holmdel, New Jersey. Since September 2000 he has been a member of the Data Analysis Research Department at Avaya Inc. He is currently a Distinguished Member of Technical Staff. His research interests include high-performance broadband packet switching architectures, local and metropolitan area network architectures, wireless communications networks, and multiuser lightwave communication networks.

REINETTE GROBLER (rg@purros.poly.edu) received B.Sc. and B.Sc. (hons.) degrees in computer science from the University of Pretoria, South Africa in 1996 and 1997 respectively. She is currently a research visitor at Polytechnic University, Brooklyn, New York, working toward an M.Sc degree in computer science at the University of Pretoria, South Africa. Her research interests include network protocols, simulation, and performance evaluation.

By implementing a signaling protocol in hardware, we propose to support call setups and releases for individual file transfers at high call handling rates and with low call setup delays.