

Control Architecture in Optical Burst-Switched WDM Networks

Yijun Xiong, *Member, IEEE*, Marc Vandenhoute, *Member, IEEE*, and Hakki C. Cankaya, *Member, IEEE*

Abstract—Optical burst switching (OBS) is a promising solution for building terabit optical routers and realizing IP over WDM. In this paper, we describe the basic concept of OBS and present a general architecture of optical core routers and electronic edge routers in the OBS network. The key design issues related to the OBS are also discussed, namely, burst assembly (burstification), channel scheduling, burst offset-time management, and some dimensioning rules. A nonperiodic time-interval burst assembly mechanism is described. A class of data channel scheduling algorithms with void filling is proposed for optical routers using fiber delay line buffer. The LAUC-VF (*Latest Available Unused Channel with Void Filling*) channel scheduling algorithm is studied in detail. Initial results on the burst traffic characteristics and on the performance of optical routers in the OBS network with self-similar traffic as inputs are reported in the paper.

Index Terms—Burst switching, channel scheduling, IP/WDM, optical routers.

I. INTRODUCTION

THE RAPIDLY-GROWING Internet is driving the demands for higher transmission capacity and high-speed IP (Internet protocol) routers at an unprecedented rate. The advances in dense wavelength division multiplexing (D-WDM) technology have made it possible to exploit the huge potential bandwidth of optical fibers (well exceeding 10 Tb/s per fiber). Currently, the D-WDM technology can already achieve 80–120 wavelengths per fiber with total transmission capacity of up to 400 Gb/s [22]. It is expected that using D-WDM, the transmission capacity per fiber may exceed 1 Tb/s in the near future. With the deployment of D-WDM technology in existing optical transport/backbone networks to meet the bandwidth requirement of Internet traffic, the routers are likely to be the bottleneck of future Internet backbones.

The past several years have seen great efforts in building hardware-based high-speed electronic IP routers. IP routers with capacity up to a few hundred gigabits per second are available now. However, there is still a serious mismatch between transmission capacity of WDM fibers and the switching capacity of electronic IP routers. A current IP router uses a crossbar switch or shared memory architecture at its backplane. To build a large IP router with capacity of 1 Tb/s or beyond, multistage interconnection network architectures will be used, which can be realized either all electronically or with the introduction of optical

switches/cross-connects in the middle stage(s). The latter approach could substantially reduce the physical size and enhance the robustness of the terabit IP router. However, in both approaches, the line cards, which typically provide SONET/SDH or gigabit Ethernet interfaces and packet forwarding function, will increasingly become a dominant factor in the total cost of the router as the router size increases.

The third approach, which is the focus of this paper, is to build terabit optical packet switches/routers and to transmit IP packets directly over WDM links on an end-to-end transparent optical path (i.e., without O/E and E/O conversions at intermediate nodes). Studies on optical packet switches/routers can be found, for instance, in [1]–[5]. This approach will avoid some functionality redundancy in intermediate layers like SONET/SDH and ATM. IP over WDM is considered a promising solution for the next generation Internet since it has fewer intermediate layers and can make better use of advanced optical technologies. Optical routers will also have better scalability (in terms of switching capacity) than electronic routers. Although line cards are largely eliminated in the optical routers, the cost of an optical router will heavily depend on the level of integration of the essential optical switching components (e.g., semiconductor optical amplifiers).

As the processing of IP packets in the optical domain is still not yet practical, the optical router control system is implemented electronically. Due to very high transmission capacity of WDM links, e.g., a single link with 32 WDM channels of 10 Gb/s each has the total transmission capacity of 320 Gb/s, the main constraint of directly switching IP packets in the optical router could be the processing and control capacity of electronic systems. An IP packet of 44 bytes lasts only 35.2 ns on a 10 Gb/s WDM channel. To reduce the burden of electronic devices (potential bottleneck) which control the configuration of an optical switching fabric and consequently increase the router throughput, the switching granularity must be larger than IP packets. This consideration leads to the concept of “burst switching” where several IP packets with the same destination and some common attributes like quality of service (QoS) are assembled into a burst and are forwarded through the network as one entity. The initial work on optical burst switching (OBS) was reported in [6]–[8].

The burst switching concept was first introduced in [10] and [11], mainly for integrated transfer of voice and data over TDM (time-division multiplexing) links. The major difference between packet and burst switching is that in packet switching, packets are transmitted at a full link speed, while in the burst switching, bursts are transmitted only at a channel speed (e.g., 64 kb/s) of a TDM link [11]. Further, the burst length could

Manuscript received October 25, 1999; revised May 15, 2000.

The authors are with the Alcatel Corporate Research Center, Richardson, TX 75081-1936 USA (e-mail: {yijun.xiong, marc.vandenhoute, candan.cankaya}@usa.alcatel.com).

Publisher Item Identifier S 0733-8716(00)09025-9.

be arbitrarily long and two special bit patterns are required to indicate the start and end of each burst. A similar scenario also arises in optical WDM networks where packets/bursts can be transmitted only at the channel speed of a WDM link. Different from the burst switching described in [11], each burst in the OBS consists of a header and a payload. The information of the payload length is carried in the header. Different from conventional store-and-forward packet switching and the burst switching in [11], the OBS uses *separate* wavelengths/channels to transmit the burst payload and its header. The burst payload is also called *data burst*, and the burst header is called *burst header packet* (BHP) in this paper.

In an OBS network, packets are assembled into bursts at network ingress and disassembled back into packets at network egress. An intrinsic feature of the OBS is the physical separation of transmission and switching of burst payloads and their headers, which helps to facilitate the electronic processing of headers at optical core routers and provide end-to-end transparent optical paths for transporting burst payloads. The OBS network can be envisioned as two coupled overlay networks: a pure optical network transferring data bursts, and a hybrid control network transferring BHPs. The control network is just a packet-switched network, which controls the routing of data bursts in the optical network based on the information carried in their BHPs. It is expected that the above separation will lead to a better synergy of both very mature electronic technologies and advanced optical technologies.

The focus of this paper is on the OBS protocols and the design of the control network. The basic concept of OBS is described in Section II along with the discussion on possible BHP and data burst formats. The general architecture of optical core routers is presented in Section III with detailed description on the switch control unit (a node in the control network). Different from [7], a conventional IP router instead of an ATM switch is used in the switch control unit. A class of data channel scheduling algorithms with void filling is proposed for optical routers with fiber delay line (FDL) buffers. The initial work on void filling channel scheduling algorithms can be found in [16]. The functional architecture of electronic edge routers is given in Section IV, and a burst assembly mechanism is proposed to assemble packets into bursts. Some fundamental issues in the OBS are also discussed in Sections III and IV. The burst traffic characteristics and the performance of optical routers in the OBS network are studied in Section V. Unlike most of the previous work, we take into account the burstification at edge routers and the electronic control at core routers in obtaining the optical router performance via computer simulation. Some further discussions are given in Section VI.

II. OPTICAL BURST SWITCHING

To circumvent potential bottlenecks of electronic processing in optical packet-type WDM networks, the basic data block to be transferred is a super packet, called *burst*, which is a collection of data packets having the same network egress address and some common attributes, like QoS requirements. A block diagram of an optical burst-switched (OBS) network is shown in Fig. 1, which consists of optical core routers and electronic

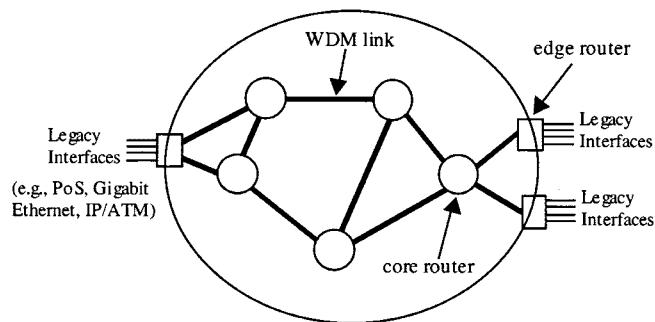


Fig. 1. An optical burst-switched network.

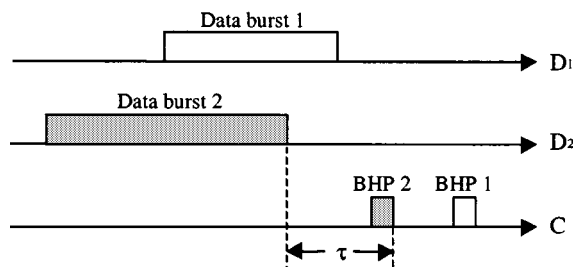


Fig. 2. Transmission of data bursts and their headers (BHPs) on a WDM link.

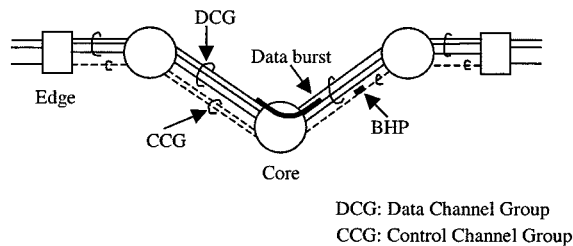


Fig. 3. Illustration of burst transmission in an OBS network.

edge routers connected by WDM links. Packets are assembled into bursts at network ingress, which are then routed through the OBS network and disassembled back into packets at network egress to be forwarded to their next hops (e.g., conventional IP routers). Edge routers provide burst assembly/disassembly functions and legacy interfaces (e.g., gigabit Ethernet, packet over SONET (PoS), IP/ATM, etc.). A core router is mainly composed of an optical switching matrix and a switch control unit (SCU).

A burst consists of a burst header and a burst payload. The burst payload is also called *data burst* in this paper. For the optical burst switching (OBS) considered here, a data burst (payload) and its header are transmitted separately on different wavelengths/channels with the burst header slightly ahead in time (see Fig. 2), and are switched in optical and electronic domains, respectively, at each core router they traverse. The burst header contains all the necessary routing information to be used by the switch control unit (SCU) at each hop to configure the optical switching matrix to switch the data burst optically (see Fig. 3). The separate transmission and switching of data bursts and their headers will help to facilitate the electronic processing of headers and lower the optoelectronic processing capacity required at core routers. Further, it can provide ingress-to-egress transparent optical paths for transporting data

bursts. As the burst header is sent in the form of a packet, it is called burst header packet (BHP) hereafter. Similar to packet switching, both connectionless and connection-oriented burst forwarding could be used in the OBS.

Throughout the paper, we use *channel* to represent a certain unidirectional transmission capacity (in bits per second) between two adjacent routers. A channel may consist of one wavelength or a portion of a wavelength, in case of time-division or code-division multiplexing. Channels carrying data bursts are called *data channels*, and channels carrying BHPs and other control packets are called *control channels* (see Fig. 3). Control packets are used to exchange routing and network information. A *channel group* is a set of channels with a common type and node adjacency. A WDM link in Fig. 1 represents a total transmission capacity between two routers, which usually consists of a *data channel group* (DCG) and a *control channel group* (CCG) in each direction. The channels of a DCG as well as its corresponding CCG could be physically carried on the same fiber or on different fibers. In the following, we use channel and wavelength interchangeably.

An example of the transmission of bursts on a WDM link is shown in Fig. 2, where the WDM link has one DCG composed of two channels and one CCG composed of only one channel. There is an offset time τ between a data burst and its BHP. The initial value τ_0 of the burst offset-time τ is set by an ingress edge router, which may be the same for all bursts or may be different from burst to burst. The function of the burst offset-time depends on the design of optical core routers. For optical core routers using input FDLs (fiber delay lines) to delay the arrivals of data bursts to the optical switching matrix, thus allowing the SCU to have sufficient time to process their BHPs, the main function of the offset time is to resolve BHP contentions on outgoing CCGs of optical core routers [7]. For optical core routers without input FDLs, the offset time should also allow the SCU at each hop along the path to have enough time to process the BHP before its associated data burst arrives. In the latter case, the burst offset-time would be proportional to the number of hops the burst will traverse in the OBS network [6], [8], and is much larger than the offset time in the former case. In both cases, the traffic condition in the network should be taken into account in choosing the offset time. The burst offset-time could also be adjusted to support QoS [12], and may play an important role in traffic scheduling/management for optical core routers without buffer or with buffer of very limited storage capacity.

To simplify the design of the SCU, in particular, the channel scheduling, optical core routers with input FDLs are considered in this paper. To have the burst offset-time well under control within the OBS network, at each hop the burst traverses, the core router tries to “resynchronize” each BHP and its associated data burst by keeping the offset time τ as close as possible to τ_0 , but no less than τ_{\min} . The typical value of τ_{\min} is zero, meaning a BHP should be sent out no later than its associated data burst. Due to the input FDLs at core routers, it is not always necessary to restrict τ_{\min} to nonnegative values, as a BHP may be behind the data burst at one node but could catch up at the next node.

An example of the data burst format is shown in Fig. 4. Each packet is delineated within the actual payload by a frame header (H). The header of the actual payload includes payload type

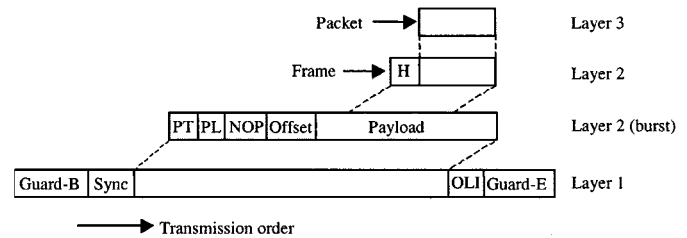


Fig. 4. An example of the data burst format at layers 2 and 1.

(PT), payload length (PL), number of packets (NOP), and the offset of padding. PT is an option indicating the type of data packets in the data burst. PL indicates the length of the payload in bytes. NOP specifies the number of packets in the payload. The offset indicates the first byte of padding. Padding may be required if a minimum burst length is imposed. In Fig. 4, the synchronization pattern in layer 1 is used to synchronize the optical receiver at the egress edge router. The guard band at the beginning (preamble) and end (postamble) of a data burst help to overcome the uncertainty of data burst arrival and data burst duration due to clock drifts between nodes, the delay variation in different wavelengths, mismatch between data burst arrival time and slotted optical switching matrix configuration time, and nondeterministic optical switching matrix configuration times. Other optical layer information (OLI) such as performance monitoring and forward error correction could also be included.

Like the packet header in conventional packet-switched networks, the BHP contains the necessary routing information to be used by core routers to route the associated data burst hop by hop to its destination edge router. Apart from the routing information carried by the conventional packet header, e.g., in IPv4, IPv6, or MPLS-like [19], the BHP contains OBS specific information as its payload which includes burst offset-time, data burst duration/length, data channel carrying the burst, the bit rate at which the data burst is sent, and QoS, among others. Various layer 1 (L1) and layer 2 (L2) technologies can be used for the control channels. One example is Packet over SONET [20].

Except for the separate transmission of headers and payloads and being switched in different domains, there is no fundamental difference between packet switching and the OBS. However, in the OBS, a burst header must explicitly reserve the switching resources in advance at each hop along the path for its burst payload, while in store-and-forward packet switching, the reservation of switching resources is made implicitly, i.e., when a packet is sent out from an electronic buffer.

The link utilization of the OBS network will largely depend on the number of channels dedicated to transmitting BHPs (as well as other control packets) and the guards in each data burst. Consider a WDM link having K channels with k control channels and $K - k$ data channels, $1 \leq k < K$. Suppose the data channel rate is R Gb/s and the control channel rate is r Gb/s. The maximum link utilization $\eta = (K - k)R / [(K - k)R + kr]$. For $K = 32$, $k = 2$, and $R = r$, $\eta = 0.9375$. As a data burst can be sent out on a data channel only if its BHP can be sent out on a control channel, there is a minimum requirement for the average data burst length in order to prevent congestion on control channels [9]. Since we will often deal with time domain issues in the OBS, it is convenient to use time duration instead of bytes

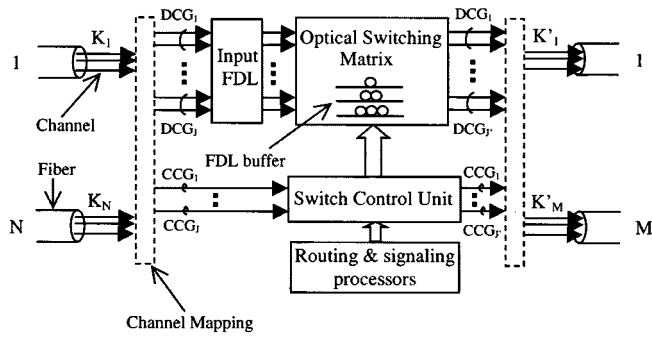


Fig. 5. A general architecture of optical routers.

to represent the length of a data burst. Without loss of generality, the basic time unit is assumed to be $1 \mu\text{s}$. Let L_b be the average duration of a data burst (or $R \cdot L_b$ kbits in length) and L_h be the average duration of a BHP (or $r \cdot L_h$ kbits in length). Consider that both control and data channels are fully loaded. Under this situation, the maximum average BHP transmission rate is k/L_h BHPs per microsecond, and the maximum average burst transmission rate is $(K - k)/L_b$ data bursts per microsecond. Since $(K - k)/L_b \leq k/L_h$, we have

$$L_b \geq (K - k)L_h/k. \quad (1)$$

For example, if $K = 32$, $k = 2$, and $L_h = 0.0512 \mu\text{s}$, the minimum average data burst duration is $0.768 \mu\text{s}$ or 0.96 kbytes in length when $R = 10$ Gb/s. Consider the guard period in Fig. 4, say $\theta \mu\text{s}$, for each data burst. The average data burst duration that actually carries user data would be only $L_b - \theta$ and the burst overhead is θ/L_b . To have the burst overhead no more than ε , it requires

$$L_b \geq \theta/\varepsilon. \quad (2)$$

For $\theta = 0.1 \mu\text{s}$ and $\varepsilon = 0.1$, $L_b = 1 \mu\text{s}$. If $\varepsilon = 0.05$, $L_b = 2 \mu\text{s}$. Inequalities (1) and (2) will together determine the minimum average data burst length.

III. OPTICAL CORE ROUTERS

A. General Architecture

The general architecture of an $N \times M$ optical core router is shown in Fig. 5, which mainly consists of input FDLs (fiber delay lines), an optical switching matrix, a switch control unit (SCU), and routing and signaling processors. Data channels are connected to the optical switching matrix and control channels are terminated at the SCU. Channel mapping logically decouples the channels from physical fibers wavelengths. The (fixed) input FDLs, if provided, are used to delay the arriving data bursts, thus allowing the SCU to have enough time to process the associated BHPs. Data bursts still remain in the form of optical signals in the core routers. The optical buffers of FDLs are used to resolve data burst contentions on outgoing DCGs (data channel groups). The use of electronic buffers instead of FDL optical buffers was considered in [7]. Note that there are J incoming DCGs and J' outgoing DCGs in Fig. 5. A typical example of the general architecture is a symmetric router with N input and N output fibers, where each fiber has one DCG of $(K - k)$ channels and one CCG (control channel group) of k channels.

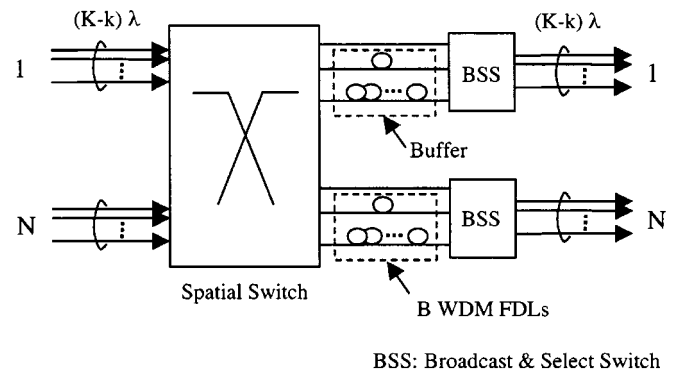


Fig. 6. Block diagram of a nonblocking (symmetric) optical switching matrix.

Various optical switching matrices, e.g., the broadcast-and-select type switch described in [4] and the switching fabrics proposed in [1]–[3], could be used in Fig. 5. Advanced optical technologies, implementation complexity, cost, and switch performance (e.g., burst loss ratio) will certainly have impact on the design of the optical switching matrix. Here we consider an ideal nonblocking optical switching matrix with output queueing. A block diagram of an $N \times N$ nonblocking optical switching matrix is given in Fig. 6 where the spatial switch is able to switch a data burst from any incoming wavelength/channel to any FDL as long as it does not overlap with other data bursts. Each optical buffer has B WDM FDLs with i th FDL being able to delay Q_i time, $1 \leq i \leq B$, and it is assumed that $Q_1 < Q_2 < \dots < Q_B$. Note that an FDL in Fig. 6 has $(K - k)$ wavelengths. By default there is always an FDL with zero delay time, denoted by 0 with $Q_0 = 0$. An example of the optical switching matrix is shown in Fig. 7 where $n = K - k$.

The function of the SCU in Fig. 5 is similar to a conventional electronic router. The routing processor runs routing and other control protocols for the whole OBS network. It creates and maintains a routing table and computes the forwarding table for the SCU. Forwarding can be connectionless or connection-oriented (prior path establishment through signaling). After forwarding table lookup, the SCU decides on which outgoing DCG and CCG to forward each arriving data burst and its BHP. If there are free data and control channels available from these groups, either when the data burst arrives to the optical switching matrix or after some delay in an FDL buffer, the SCU will then select the FDL of the optical buffer and configure the optical switching matrix to let the data burst pass through. Otherwise, the data burst is dropped. In arranging the transfer of a data burst and its corresponding BHP in the optical switching matrix and SCU, respectively, the SCU tries to resynchronize the data burst and the BHP by keeping the offset time τ as close as possible to τ_0 .

If a data burst enters the optical switching matrix before its BHP has been processed (this phenomenon is called *early burst arrivals*), the burst is simply “dropped.” This is because data bursts are optical analog signals. If no path is set up when a data burst enters the optical switching matrix, it is lost. Since a BHP and its data burst are switched in the SCU and the optical switching matrix, respectively, the delay Δ introduced by the input FDL should be properly engineered such that under the normal traffic condition data bursts are rarely dropped due to early arrivals.

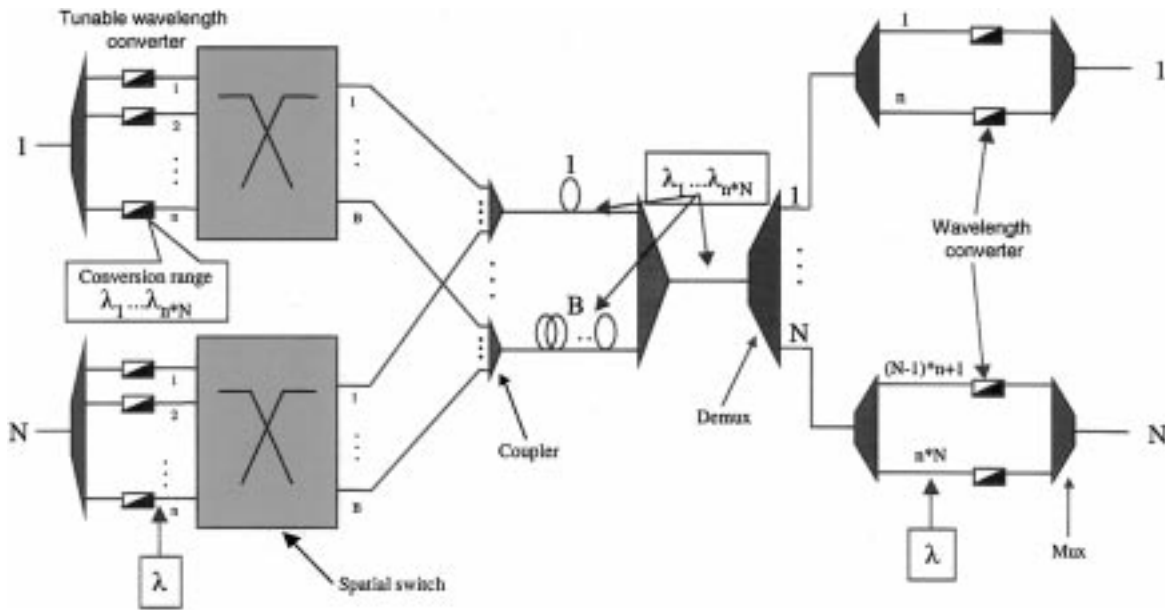


Fig. 7. An example of the optical switching matrix.

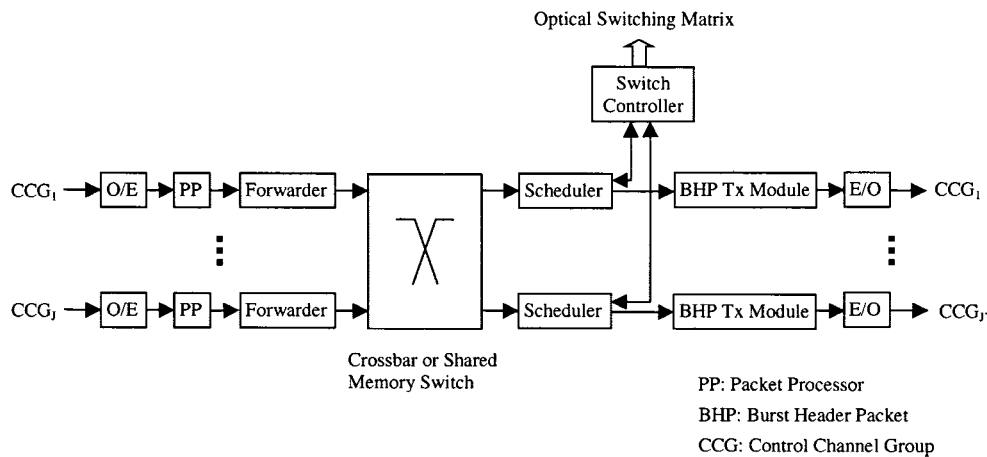


Fig. 8. Block diagram of the switch control unit (centralized control).

Since we mainly deal with the control architecture of the OBS network, more details on the SCU are described below. The block diagram of a SCU is given in Fig. 8. Depending on the optical switching matrix, the SCU can have either a centralized configuration as shown in Fig. 8 or a distributed configuration. In a distributed configuration, each scheduler has its own switch controller. Distributed configuration could be applied to the broadcast-and-select type switches [4]. Here we focus on the description of centralized configuration. The functionality of each building block in Fig. 8 is detailed below.

The packet processor (PP) performs L1 and L2 decapsulation functions and attaches a time-stamp to each arriving BHP, which records the arrival time of the associated data burst to the optical switching matrix. The time-stamp is the sum of the BHP arrival time, the burst offset-time τ carried by the BHP, and the delay Δ of input FDL. The forwarder performs the forwarding table lookup to decide on which outgoing CCG to forward the BHP. The associated data burst will be forwarded to the corresponding DCG. The mapping of logical channels to physical wavelengths and fibers is done in the forwarder. The forwarder

then simply forward BHPs across the switch in a certain order (e.g., FIFO). To reduce the switch delay, it is preferred to use a switch with output-queueing discipline. To support multicast traffic, the switch requires native multicast capability. Otherwise, copies of a multicast BHP are made in the forwarder.

The scheduler in Fig. 8 is responsible for both the scheduling the switch of the data burst on an outgoing data channel and the scheduling the transmission of its BHP on an outgoing control channel. The scheduler is optical switching matrix specific. For the nonblocking optical switching matrix in Fig. 6, there is one scheduler for each DCG and CCG pair, and each scheduler only needs to keep track of the busy/idle periods of a single outgoing DCG and an outgoing CCG. The scheduler works as follows. It first reads the time-stamp and the data burst duration information from a BHP to determine when the corresponding data burst will enter the optical switching matrix and how long the data burst will last. It then searches for an idle outgoing data channel time slot to carry the data burst, making potential use of the FDLs to delay the data burst. Once the idle outgoing data channel is found and the FDL to be used (if necessary) is de-

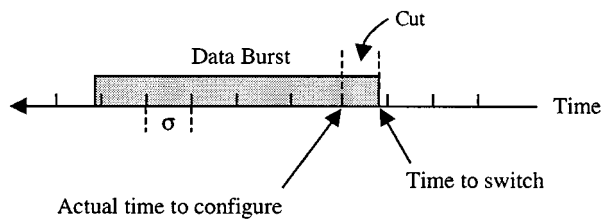


Fig. 9. Mismatch of time-to-switch and optical switching matrix configuration time

terminated, the scheduler knows the departure time of the data burst from the optical switching matrix. Subsequently, it schedules the time to send out the BHP on the outgoing CCG, trying to resynchronize the BHP with the data burst. After successfully scheduling the transfer of the data burst and its BHP, the scheduler will send the configuration information to the switch controller which will in turn configure, just-in-time, the optical switching matrix accordingly to let the data burst pass through. The configuration information includes incoming data channel identifier, outgoing data channel identifier, time to switch the data burst, duration of the data burst, and the FDL buffer identifier.

The scheduler is bidirectionally connected to the switch controller. After processing the configuration information sent by the scheduler, the switch controller sends back an acknowledgment to the scheduler. The scheduler then updates the state information of the DCG and CCG, modifies the BHP (e.g., the offset time and the data channel identifier) and passes it along with the time-to-send BHP information to the BHP transmission (Tx) module. It is now ready to process the next BHP. Some pipelining may be used here to speed up the scheduling process. The BHP transmission module sends out the BHP at the pre-specified time and performs L2 and L1 encapsulation functions. One of the reasons we need an acknowledgment from the switch controller is to limit the number of configuration requests from the schedulers. Hence, the maximum time for the switch controller to process the request can be estimated.

The switch controller configures the optical switching matrix in a time slotted fashion, say every $\sigma \mu s$. Nonslotted configuration of the optical switching matrix is difficult to implement due to the asynchronous arrivals of data bursts. If the WDM transmission system is also slotted, then the switch controller can use the same slot length and be synchronized with the slotted transmission. For the nonslotted WDM transmission system, the actual optical switching matrix configuration time is not necessarily equal to the time-to-switch from the scheduler. A small portion at the beginning of the burst could be cut as depicted in Fig. 9. However, the real data will not be cut if the guard-B in Fig. 4 is larger than σ . The new burst offset time is calculated using the time-to-switch, *not* the actual matrix configuration time, so the guard band of the data burst at next hop is still guard-B.

In the case where the required delay time for the data burst in the optical switching matrix is too long, e.g., longer than Q_B , or the BHP cannot be sent out on the outgoing CCG due to congestion or there is not enough time to process the BHP before the data burst enters the optical switching matrix, the data burst and its corresponding BHP are simply discarded.

The scheduler is the key component in the SCU. The design of the scheduler poses a new challenge, as it has to work at ex-

tremely high speed (e.g., 100–200 ns per BHP) and the gaps/voids introduced by the FDL optical buffer will greatly complicate the design. Massive parallelism and pipelining are required in hardware implementation. The core part of the scheduler is the DCG scheduling. We now describe some data channel scheduling algorithms that could be used in the scheduler.

B. Data Channel Scheduling

Data channel scheduling algorithms can be classified into two categories: without [7], [17] and with [16] void filling (VF). In this subsection, we first describe a simple scheduling algorithm without void filling, called LAUC (*latest available unscheduled channel*) algorithm [17], which is very similar to, if not exactly the same as, the Horizon algorithm proposed in [7]. We then extend it to a more sophisticated scheduling algorithm by incorporating void filling, which is called LAUC-VF (*latest available unused channel with void filling*). Some variations of these scheduling algorithms will also be discussed.

It is assumed that the optical buffer has B FDLs with i th FDL being able to delay Q_i time (see Fig. 6), $1 \leq i \leq B$. For FDL 0, its delay time $Q_0 = 0$. To simplify the description, we assume $Q_i = i \cdot D$ where D is a given time unit. It is further assumed that the switching latency of the spatial switch and broadcast-and-select switches in Fig. 6 are negligible, hence the data burst arrival time to the optical switching matrix is equal to its departure time if FDL 0 is used.

1) *LAUC Algorithm:* In the LAUC algorithm, only one real value—the unscheduled time (future available time)—is maintained for each data channel of an outgoing DCG. For a DCG with $K - k$ data channels, let t_j be the unscheduled time of j th channel, $j = 1, 2, \dots, K - k$. Because the arrival order of BHPs is not necessarily the arrival order of their data bursts to the optical router due to the variable offset-time and the queueing delay in the SCU, the basic idea of the LAUC algorithm is to minimize gaps/voids by selecting the latest available unscheduled data channel for each arriving data burst. Given the arrival time t of a data burst with duration L to the optical switching matrix, the scheduler first finds the outgoing data channels that have not yet been scheduled at time t . If there is at least one such channel, the scheduler selects the latest available channel, i.e., the channel having the smallest gap between t and the end of last data burst just before t , to carry the arriving data burst. The selected channel's unscheduled time (i.e., the future available time) is then updated to $t + L$. For example, in Fig. 10(a) data channels 2 and 3 are unscheduled channels at time t , and D_2 is selected to carry the arriving data burst as $t - t_2 < t - t_3$.

If all channels were already scheduled at time t , the arriving data burst has to be delayed by a multiple of FDL units, say i units, until at least one unscheduled data channel is found. If $1 \leq i \leq B$ (B is the maximum number of FDL units), the scheduler will select the latest available channel to transmit the data burst and update the channel's unscheduled time to $t + i \cdot D + L$. If $i > B$, the arriving burst is simply discarded. In Fig. 10(b), all data channels were already scheduled at time t but channels 1 and 3 are unscheduled at $t + D$. So the arriving data burst will be delayed for one FDL unit and channel D_3 is selected to carry the data burst. Note that voids could be generated due to different data burst arrival times [Fig. 10(a)] or FDL buffer increments

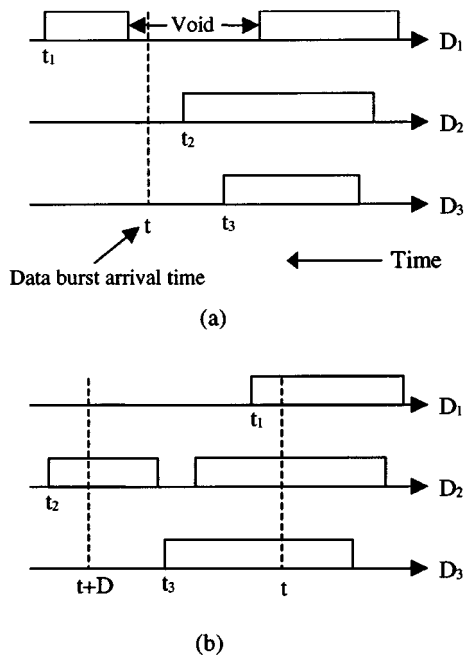


Fig. 10. Illustration of LAUC algorithm, (a) channel 2 is selected, (b) channel 3 is chosen.

[Fig. 10(b)]. Obviously, the larger the FDL unit D , the bigger the void could be.

Simplicity and ease of implementation are the main advantages of the LAUC algorithm as the scheduler needs only to remember one value—the unscheduled time—for each data channel. The simplicity is very important in extremely high speed environments. The drawback of the LAUC algorithm is the inefficient use of data channels as the gaps/voids between data bursts are not utilized. The storage capacity of the FDL buffer is determined by not only the number of FDLs but also the length of each FDL. Obviously, the larger the FDL unit D , the bigger the void introduced which makes the LAUC algorithm less effective, as a consequence causing higher burst loss ratio (see Fig. 19 in Section V). To solve this problem, sophisticated scheduling algorithms with void filling may be considered.

A simpler version of the LAUC algorithm is called the FF (*first fit*) algorithm, where the data channels are searched in a given order, e.g., fixed or round-robin, and the first eligible channel found, instead of the latest available unscheduled channel, will be used to carry the data burst. The comparison of the LAUC and FF algorithms will be given in Section V.

2) *LAUC-VF Algorithm*: The void/gap between the two data bursts in data channel D_1 of Fig. 10(a) is unused channel capacity. This algorithm is similar to the LAUC algorithm except the voids can be filled by new arriving data bursts. The basic idea of the LAUC-VF algorithm is to minimize voids by selecting the latest available *unused* data channel for each arriving data burst. Unscheduled data channels are just a special case of unused data channels. Given the arrival time t of a data burst with duration L to the optical switching matrix, the scheduler first finds the outgoing data channels that are available for the time period of $(t, t + L)$. If there is at least one such data channel, the scheduler selects the latest available data channel, i.e., the channel having the smallest gap between t and the end of last data burst just before t .

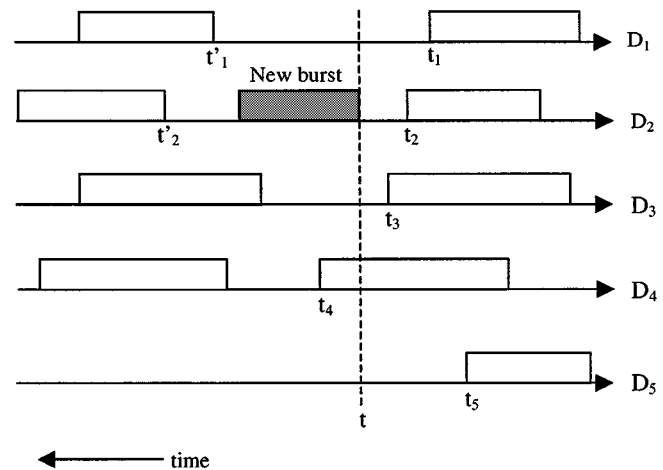


Fig. 11. Illustration of LAUC-VF algorithm.

Fig. 11 shows an illustration of the LAUC-VF algorithm. In Fig. 11, the DCG has 5 data channels where D_1 , D_2 , and D_5 are eligible unused data channels at t for carrying the data burst. However, data channels D_3 and D_4 are ineligible at t because the void is too small on D_3 for the data burst and D_4 is busy at t . Data channel D_2 is chosen to carry the data burst as $t - t_2 < t - t_1 < t - t_5$. If all the data channels are ineligible at time t , the scheduler will then try to find the outgoing data channels that are eligible at time $t + D$ [i.e., available for the time period of $(t + D, t + D + L)$], and so on. If no data channels are found eligible up to time $t + B \cdot D$ [i.e., for the time period of $(t + B \cdot D, t + B \cdot D + L)$], the arriving data burst and the corresponding BHP are dropped. Note that $B \cdot D$ constitutes the longest time the data burst can be buffered (delayed).

The formal description of the LAUC-VF algorithm is presented below. Let $Ch_Search(x)$ be a function which searches for the eligible latest available unused channel at time x and returns the selected outgoing data channel to carry the data burst if found, otherwise return value -1 . Let t be the data burst arrival time to the optical switching matrix and let j be the outgoing data channel selected to carry the data burst.

```

Begin {LAUC-VF algorithm}
  Step 1:  $i = 0$ ;  $x = t$ ;
  Step 2:  $j = Ch\_Search(x)$ ;
           if ( $j \neq -1$ )
             {report the selected data
              channel  $j$  and the selected FDL
               $i$ ; stop;}
           else
             {
                $i = i + 1$ ;
               if ( $i > B$ )
                 {report failure in finding
                  an outgoing data channel
                  and stop;}
               else
                 { $x = t + Q_i$ , goto Step 2;}
             }
End {LAUC-VF algorithm}

```

Note that in the LAUC-VF algorithm, Q_i is not further specified, $1 \leq i \leq B$. An exhaustive search is used in the above algorithm, i.e., considering all possible delay times provided by the FDL buffer (starting from Q_1 to Q_B) when necessary in searching for an available outgoing data channel. Due to the stringent time constraint, a limited search may be preferred when B is large (e.g., $B \geq 16$), which only uses a subset of $\{Q_1, Q_2, \dots, Q_B\}$ in searching for an available outgoing data channel.

For a given time t , the data channels can be classified into unscheduled channels where no data bursts are scheduled after t (e.g., D_5 in Fig. 11) and scheduled channels in which data bursts are already scheduled after t (e.g., D_1, D_2, D_3 , and D_4). The above LAUC-VF algorithm does not distinguish between scheduled and unscheduled data channels. Some variations of the LAUC-VF algorithm are listed below.

- 1) The data channels are searched in the order of scheduled and unscheduled channels for a given time instant.
- 2) The data channels are searched in the order of scheduled and unscheduled channels. For eligible scheduled channels, the channel with minimum gap is chosen (e.g., the channel selected is D_1 if $t'_1 - t_1 < t'_2 - t_2$). The LAUC principle still applies for the eligible unscheduled channels.
- 3) The data channels are still searched in the order of scheduled and unscheduled channels. The first eligible scheduled channel is chosen. If all scheduled channels are ineligible, then the first eligible unscheduled channel is chosen. The round-robin could be used for each type of channels.
- 4) The data channels are searched in an order, either fixed or round-robin. The first eligible channel is chosen (in this case, it is channel D_1 in Fig. 11). This algorithm is called FF-VF.

The LAUC algorithm is a special case of the LAUC-VF algorithm by restricting $Ch_Search(x)$ function to unscheduled data channels. In the above, we actually present a class of scheduling algorithms that could be used in optical routers with FDL buffer. The hardware implementations of these algorithms are described in [18]. Special-purpose parallel processing architecture is used in the design of the scheduler to meet the stringent real-time requirement. Specifically, we use associative memory to store the state information of channels, and use associative processor arrays to implement void/gap search and channel state information update. The implementation complexity will depend on the DCG size, the BHP delay in the SCU, the maximum data burst length, the FDL buffer, and the data burst characteristics.

C. Dimensioning Issues

As the transmission and switching of burst headers and their payloads are physically separate, bursts can get lost in an optical router due to the congestion in outgoing DCGs, the congestion in outgoing CCGs, and BHP loss or excessive delay in the SCU. The latter two could be avoided via proper dimensioning of input FDLs, the CCG and the SCU, and regulating the burst arrival rate to the SCU.

There are two dimensioning issues in the SCU of Fig. 8: 1) buffer sizes in the forwarder, the switch, and the scheduler; 2) BHP delay from entering the SCU until being processed by the BHP transmission (Tx) module. Ideally, burst loss should not be caused by the SCU. So the buffer sizes have to be dimensioned large enough that there is no BHP loss due to buffer overflow. Apart from the fixed delays, queues in the forwarders, the switch, and the schedulers of the SCU will introduce some variable delays. The delay Δ of the input FDL in Fig. 5 should be dimensioned sufficiently long such that the probability of the BHP delay larger than Δ is extremely small, say 10^{-11} . The switch in the SCU should not be heavily loaded in order to reduce the switching latency.

The size of a CCG will depend on the capacity of its corresponding DCG and the traffic volume of control packets to be sent on the CCG. Given the CCG size, the initial burst offset-time τ_0 is a very important parameter in the OBS. Small τ_0 may cause unnecessary burst loss, while large τ_0 may lead to inefficient use of data channels and increase the control complexity (width of scheduling window) in the SCU. It is expected that for a given traffic load, τ_0 will increase with the router size as well as the number of hops. The reason is that when the router size is large, the data burst contention for an outgoing DCG will be higher, resulting in more variation in the burst offset-time.

Next, let us look at the impact of the processing speed of the forwarder, the scheduler, and the switch controller on the burst arrival rate to an $N \times N$ optical router, where each fiber has a DCG of $K - k$ channels and a CCG of k channels. Suppose the BHP processing times for the forwarder, the scheduler, and the switch controller are δ_f , δ_s , and δ_c , respectively, all in microseconds. Let λ be the average burst arrival rate (bursts/ μ s) per fiber. In Fig. 8, the forwarder, the scheduler, and the switch controller can each be modeled as a single-server queueing system. For a stable queueing system, the service rate must be larger than the arrival rate. We have $\lambda < 1/\delta_f$, $\lambda < 1/(\delta_s + X)$, and $\lambda < 1/(N \cdot \delta_c)$, which leads to

$$\lambda < \min \left\{ \frac{1}{\delta_f}, \frac{1}{\delta_s + X}, \frac{1}{N \cdot \delta_c} \right\} \quad (3)$$

where $X(\delta_c \leq X \leq N \cdot \delta_c)$ is the average time spent in the switch controller. For distributed control, $X = 0$ and N is replaced by 1 in the above inequality. Let ρ be the data channel utilization. Given λ , the average burst duration (in μ s) can be expressed as

$$L_b = \frac{\rho(K - k)}{\lambda}. \quad (4)$$

Obviously, the average burst duration is proportional to the number of data channels per fiber. For $N = 8$, $K = 32$, $k = 2$, $\rho = 0.8$, $\delta_f = 0.1 \mu$ s, $\delta_s = 0.1 \mu$ s, and $\delta_c = 0.025 \mu$ s, we have $\lambda < 5$ bursts/ μ s, which leads to $L_b > 4.8 \mu$ s (or 6 kbytes if the data channel rate $R = 10$ Gb/s). However, if $\delta_c = 0.1 \mu$ s, $L_b > 19.2 \mu$ s (or 24 kbytes if $R = 10$ Gb/s). Note that inequalities (1) and (2) are from the burst transmission viewpoint, while (3) and (4) consider the switching aspect. The burst arrival rate is also closely related to the burst assembly mechanism used at edge routers.

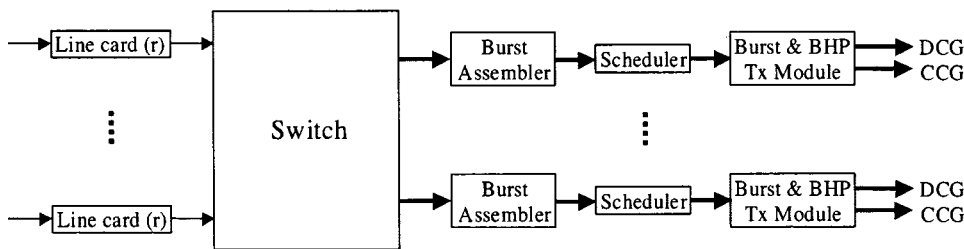


Fig. 12. Functional architecture of edge routers (sending part).

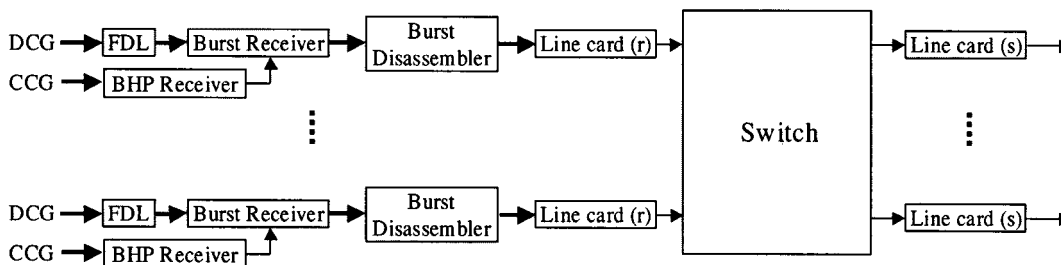


Fig. 13. Functional architecture of edge routers (receiving part).

IV. ELECTRONIC EDGE ROUTERS

A. Functional Architecture

Edge routers represent the deaggregation and transit points between an OBS network and legacy domain of any inter-networking architecture. An edge router connects multiple subnetworks running on top of legacy link layer protocols to the OBS network. The simplified functional architecture of edge routers is shown in Fig. 12 (sending part) and Fig. 13 (receiving part). The line cards and switching fabric in these two figures are from conventional routers. Each line card is decomposed into a receiving part and a sending part, denoted by line card (r) and line card (s), respectively. L1 and L2 decapsulation functions and packet forwarding, which includes routing table lookup, traffic classification, policing/shaping, etc., are performed in line card (r). Line card (s) mainly performs L2 and L1 encapsulation functions. The main function of the sending (ingress) part of edge routers is to assemble packets into bursts and forward them to the core network according to the OBS protocol. The additional function on the line card (r) in Fig. 12 is to attach an egress edge router address to a packet (assuming connectionless forwarding), which will be used later by the burst assembler.

The burst assembler in Fig. 12 assembles packets into bursts according to their egress edge router addresses and QoS requirements. For multicast traffic, burstification is performed based on multicast group addresses (other, more sophisticated, options are described and studied in [21]). The scheduler schedules the transmission of bursts in a certain order according to burst types and QoS requirements. It keeps track of the unscheduled time (i.e., the future available time) for each data channel. It also keeps track of the unscheduled time for each control channel. For a given burst, the scheduler tries to find the earliest times to send the data burst and its BHP on a data channel and a control channel, respectively. An offset time, say τ_0 , is maintained

between the BHP and its data burst. The burst and BHP transmission module is responsible for transmitting the BHP and the data burst at the prespecified times.

The receiving (egress) part of the edge router (Fig. 13) is similar to the ingress of an optical core router. An FDL is used here to allow the BHP receiver to have enough time to process the BHP and to instruct the burst receiver to receive the corresponding data burst. After the data burst is received, it is sent together with the information carried by the BHP to the burst disassembler where bursts are simply disassembled back into packets. These packets are then forwarded to their next hops in the same way as in a conventional router. Burst reordering and retransmission will be handled in the burst disassembler if required. Note that L1 functions may not be required in line card (r) in Fig. 13. Parallelism can be easily applied here, specifically, for each data channel, there will be one burst receiver, one burst disassembler, and one line card (r).

B. Burst Assembly Mechanism

Here we propose a burst assembly mechanism based on egress edge router addresses, assembly time intervals, and a maximum data burst size. For simplicity, only unicast traffic is considered. Suppose there are G destinations (egress edge router addresses) and the OBS network provides S different QoS classes, each burst assembler in Fig. 12 needs $S \cdot G$ buckets to sort arriving packets. Assume the burst assembly time of bucket i is $T_a(i)$ μ s, which could be adaptive to the traffic condition in the OBS network, $1 \leq i \leq S \cdot G$. Let the timer of bucket i be denoted by $T_c(i)$ and the burst length (in bytes) in the bucket i by $l_b(i)$. The basic idea is to start counting the time when there is a packet arrival to empty bucket i , and when either the elapsed time equals $T_a(i)$ or the number of bytes in the bucket reaches L_{\max} bytes, the maximum data burst size, a burst is assembled with length $l_b(i)$ and $l_b(i)$ is then reset to zero. The detailed procedure is given as follows.

A Nonperiodic Time-Interval Burst Assembly Mechanism:

- 1) When a packet with length of x bytes arrives to bucket i :
 - If $(l_b(i) = 0)$
 - $\{T_c(i) = 0; l_b(i) = x;\}$
 - else if $(l_b(i) + x < L_{\max})$
 - $l_b(i) = l_b(i) + x;$
 - else
 - $\{$ report the arrival of a burst with length $l_b(i);$
 - $T_c(i) = 0; l_b(i) = x;\}$.
- 2) When $T_c(i) = T_a(i)$
 - $\{$ report the arrival of a burst with length $l_b(i);$
 - $l_b(i) = 0;\}$.

In the above time-interval based burst assembly mechanism, parameter $T_a(i)$ will determine the burst arrival rate λ on a fiber. To prevent congestion in the SCU, the burst arrival rate λ has to satisfy inequality (3). Clearly, how to choose $T_a(i)$ is a fundamental issue in the OBS. Suppose there is only one burst assembly time for all destinations and QoS, i.e., $T_a(i) = T_a$, $1 \leq i \leq S \cdot G$. As the burst assembly time starts when an empty bucket receives the first packet, it is clear that $\lambda < S \cdot G / T_a$. If the switch controller is the bottleneck in the SCU, we have from inequality (3) that $\lambda(N \cdot \delta_c) < 1$ which leads to

$$T_a > N \cdot S \cdot G \cdot \delta_c. \quad (5)$$

In the above simple example, one can see that the burst assembly time increases linearly with the number of destinations and the number of QoS classes. The increase in the burst assembly time will introduce longer packet delay in edge routers. How to choose burst assembly times in the whole OBS network environment is still an open question.

V. PERFORMANCE STUDY

In this section, we study the burst traffic characteristics and the performance of optical routers via computer simulation. The simulation environment is shown in Fig. 14, where it is assumed that each WDM link has one DCG and one CCG, and K channels in total in each direction. The DCG has $K - k$ channels and the CCG has k channels. The data channel rate is R Gb/s and the control channel rate is r Gb/s. Edge routers connect legacy packet subnetworks to the OBS backbone network. To feed the data channels of a WDM fiber, an edge router needs to connect many IP routers via, for instance, OC-192 links. Our focus here is on the burst traffic characteristics after burstification at edge routers and on the performance of the first core router that connecting the edge routers. How to characterize the burst traffic on any link within the OBS network is a complicated issue which is beyond the scope of this paper.

A. Traffic Model and Burst Characteristics

The packet stream from an IP router to the edge router is modeled here by the fractional Gaussian noise (FGN) self-similar traffic [14] with Hurst parameter H . The packet length distribution used in our study is shown in Fig. 15, which reflects the fact of predominance of small packets, with peaks at the common sizes of 44, 552, 576, and 1500 bytes [13]. The average packet length in Fig. 15 is 389.5 bytes. Specifically, let l_p be the packet length in bytes. We have $\Pr[l_p = 44] = 0.5$, $\Pr[l_p =$

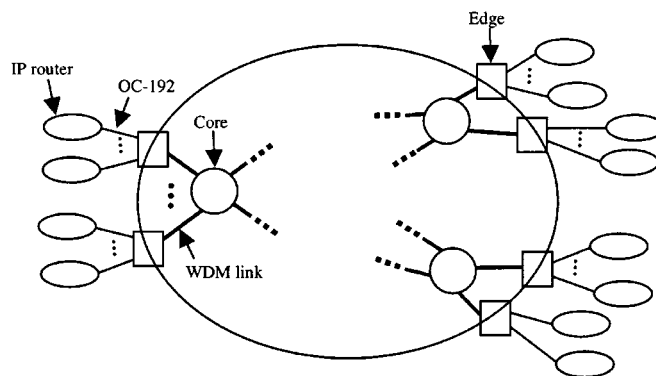


Fig. 14. The simulated OBS network.

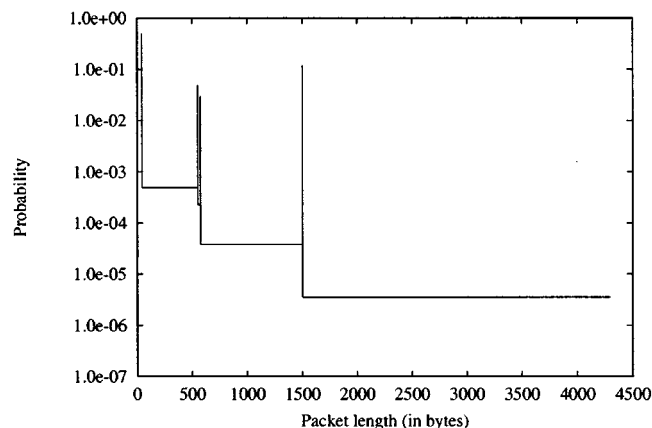


Fig. 15. Probability distribution of packet length.

$552] = 0.05$, $\Pr[l_p = 576] = 0.03$, $\Pr[l_p = 1500] = 0.12$, $\Pr[45 \leq l_p \leq 551] = 0.25$, $\Pr[553 \leq l_p \leq 575] = 0.005$, $\Pr[577 \leq l_p \leq 1499] = 0.035$, and $\Pr[1501 \leq l_p \leq 4300] = 0.01$. The mean (m) and variance (v) of number of packets in an interval are important parameters in the FGN model, and we assume $v = m^2$. As the FGN model only specifies the number of packets in a given time interval, not the arrival time of each packet, a method is developed to determine the packet arrival process. Specifically, the length of each packet is independent and identically distributed. The packets in a given time interval will arrive consecutively (rather worse case), which may be prolonged to the next time interval. Of course, by doing so, the self-similarity of the original traffic will be changed (see Fig. 17). Another traffic model to be considered is the random traffic ($H = 0.5$), where the interpacket time has an exponential distribution.

Fig. 16 shows a typical example of the burst length distribution, in which the number of destinations G is 8, $K = 16$, $k = 1$, $R = 10$ Gb/s, the data channel utilization is 0.76, $H = 0.8$, and the burst assembly time $T_a = 2$ and $8 \mu\text{s}$. An interval of $10 \mu\text{s}$ is used in the FGN model which leads to $m = 24.4$ packets. These results are not difficult to understand as the length of a burst is the sum of lengths of packets collected in a given assembly time T_a . The longer the assembly time T_a , the more likely the shape of the burst length distribution is to approach that of Gaussian distribution. Very similar results are obtained when increasing G but keeping the ratio G/T_a constant. Similar results are also observed when $H = 0.5$ [17] and when using the

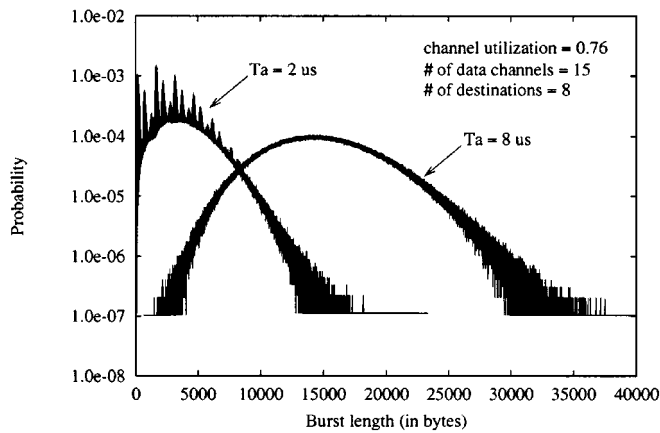


Fig. 16. Probability distribution of burst length.

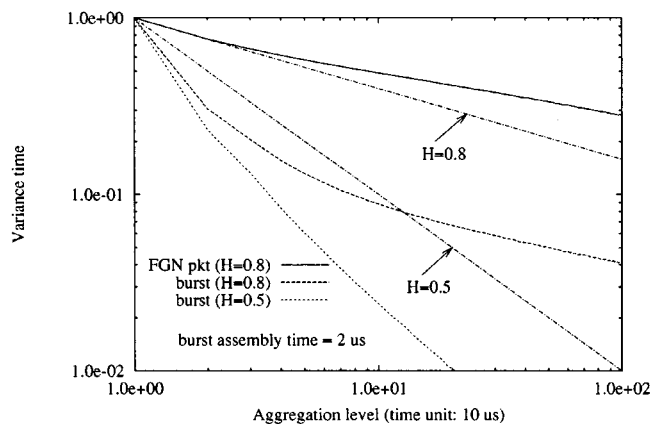


Fig. 17. Variance-time plots of packet and burst traces.

real IP packet traces from vBNS (very high performance backbone network service) [15]. The results in Fig. 16 indicate that it may not be appropriate to use an exponential type of distribution for the burst length in the OBS related performance study.

The variance-time plots of the FGN packet trace ($H = 0.8$) and the resulting burst trace are depicted in Fig. 17, using the same parameters as in Fig. 16. It shows that the Hurst parameter H of the modified FGN packet trace is a bit larger than 0.8 at high aggregation level. Although the Hurst parameter of the burst trace seems unchanged at high aggregation level (e.g., $>100 \mu s$), it is much smaller, even less than 0.5, at low aggregation levels (e.g., $<50 \mu s$). In other words, the burst trace is only asymptotically self-similar. Note that smaller H values at low aggregation level do not necessarily mean that the traffic burstiness is substantially reduced, as the burst length is much larger than the packet length, although the buffers in the burst assembler of Fig. 12 may smooth the traffic a little bit. For comparison, the variance-time plots of the random traffic packet trace ($H = 0.5$) and the related burst trace are also shown in Fig. 17. The R/S plot under the same condition is depicted in Fig. 18 for completeness.

B. Burst Loss Ratio in Core Routers

In this subsection, we study the performance of an $N \times N$ optical router with output queueing (Fig. 6), which is connected by edge routers via WDM fibers as shown in Fig. 14. The following

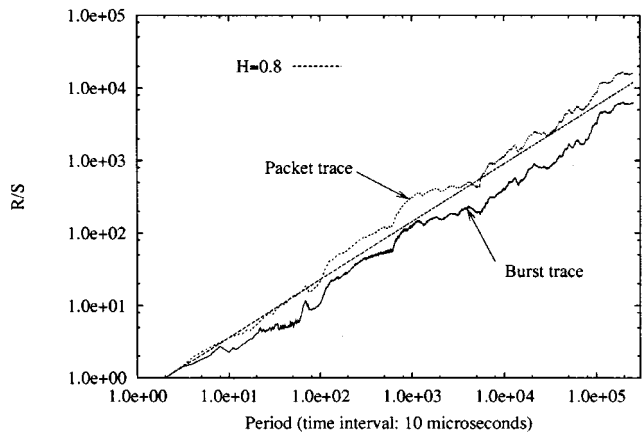
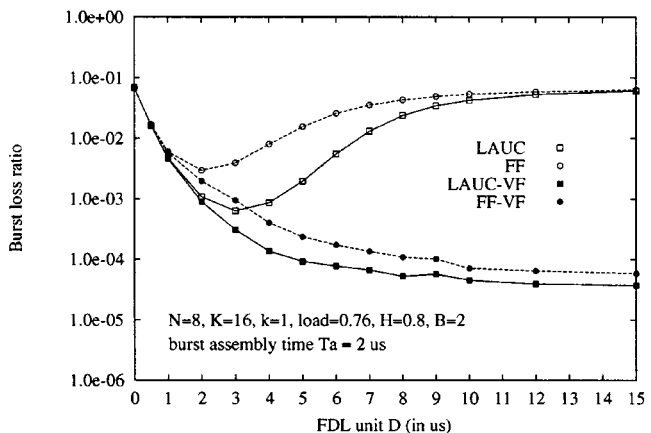
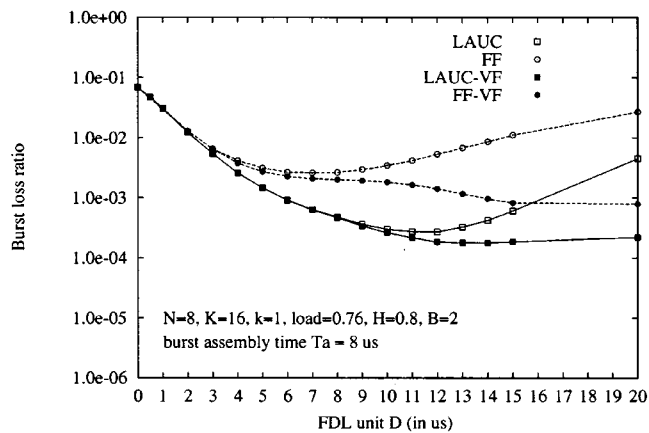


Fig. 18. R/S plots of packet and burst traces.

Fig. 19. Burst loss ratio under self-similar traffic ($T_a = 2 \mu s$).

parameters are used: $N = 8$, $K = 16$, $k = 1$, $R = r = 10$ Gb/s. It is assumed that each burst entering the core router will be routed to an outgoing port with probability $1/N$, independent of other bursts. This is an important assumption which implies that the performance of the core router will not be sensitive to G , the number of edge routers, and T_a , the burst assembly time, as long as the ratio G/T_a is kept constant. The average BHP length is 64 bytes. As data bursts and their BHPs are switched in parallel in the optical router (see Fig. 5), there are three main sources that will cause burst loss: 1) congestion in the outgoing data channels, 2) congestion in the outgoing control channel, and 3) BHP loss or excessive delay due to SCU internal congestion. The latter two could be avoided via proper dimensioning. The focus of our study here is on the burst loss due to the data channel congestion. For optical buffers using WDM fiber delay lines (FDLs), the storage capacity of an FDL buffer depends on the number of FDLs B , the length of each FDL, as well as the number of wavelengths per FDL (which is $K - k$, in our case). The FDLs are assumed to be equally spaced, i.e., $Q_i = i \cdot D$, $1 \leq i \leq B$, where D is the basic delay time unit in μs .

The burst loss ratio versus delay D is shown in Fig. 19 under $H = 0.8$, $T_a = 2 \mu s$, $B = 2$, and the channel utilization/load of 0.76. One can see that the LAUC and the FF (*First Fit*) scheduling algorithms described in Section III are quite sensitive to D . This is due to the fact that increasing D will inevitably enlarge the voids between data bursts although the total

Fig. 20. Burst loss ratio under self-similar traffic ($T_a = 8 \mu s$).

buffer storage capacity is also increased. Apparently, the LAUC algorithm is better than the FF algorithm. However, for both algorithms, when D is too long, it is equivalent to having no buffer. For $T_a = 2 \mu s$, the average burst length is 4042.6 bytes or $3.23 \mu s$ when $R = 10$ Gb/s. Using the LAUC algorithm, the lowest burst loss ratio is roughly obtained when D is equal to the average burst duration. Further, applying void filling (VF), the burst loss ratio can be substantially reduced and becomes less sensitive to the delay unit D , as D increases (comparing LAUC with LAUC-VF in Fig. 19). Of course, VF makes sense only when D is reasonably large. In this example, the difference between LAUC-VF and FF-VF is not very significant.

It is clear from Fig. 19 that channel scheduling algorithms without void filling, such as LAUC and FF, although simple, may have some potential problems as it is very difficult to know, if at all possible, the burst traffic characteristics in advance. The burst traffic characteristics are dynamic and changes with time. So it is likely that the burst assembly time intervals will be adaptive and vary with traffic types and destinations, should the LAUC algorithm be used. On the other hand, the channel scheduling algorithms with void filling, such as LAUC-VF, in spite of its high temporal and spatial complexity, can reduce the burst loss ratio and simplify the burstification processes at edge routers.

Similar results are observed when the burst assembly time $T_a = 8 \mu s$, as shown in Fig. 20. When $T_a = 8 \mu s$, the average burst length is 14980.8 bytes or $11.99 \mu s$ if $R = 10$ Gb/s. Compared to Fig. 19, the differences between LAUC and FF as well as between LAUC-VF and FF-VF are more pronounced. For LAUC-VF, the burst loss ratio under $T_a = 8 \mu s$ is much higher than the burst loss ratio when $T_a = 2 \mu s$, indicating that longer burst assembly time may potentially cause more data burst contention on the outgoing DCGs of an optical router. We also found from our simulation results that the difference between LAUC-VF and its other variations, described in Section III (e.g., minimum gap, scheduled and unscheduled channel ordering, etc.), is very small, almost negligible. The burst loss ratio under traffic load of 0.86 is shown in Fig. 21. In this case, the average burst length is 4478.19 bytes or $3.58 \mu s$ for $R = 10$ Gb/s. Compared to Fig. 19, the value of D which yields the lowest burst loss ratio seems to decrease as the traffic load increases.

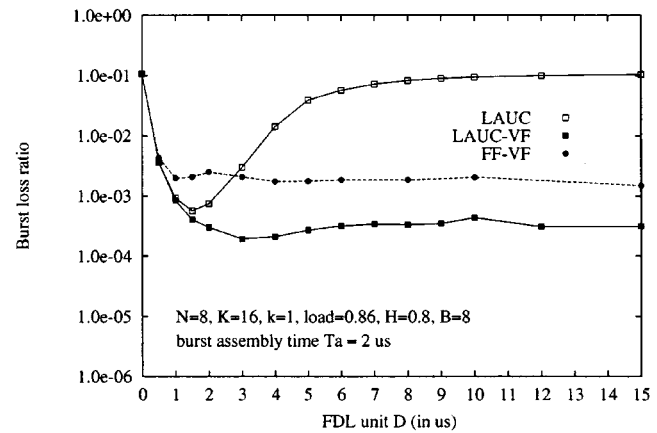
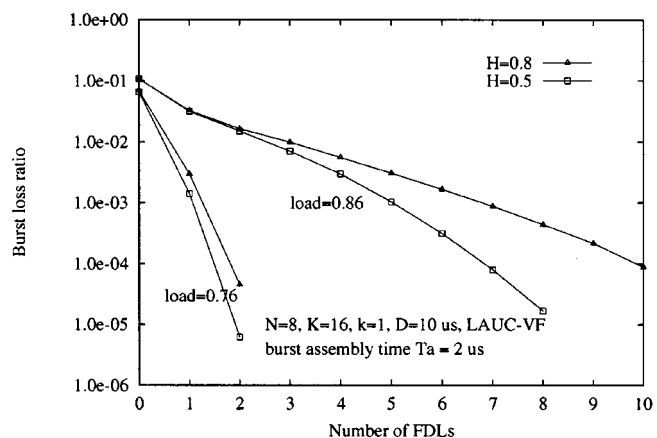


Fig. 21. Burst loss ratio under self-similar traffic (load = 0.86).

Fig. 22. Impact of buffer size B on burst loss ratio.

For higher traffic load, the difference between LAUC-VF and FF-VF appears to be more pronounced.

The burst loss ratio versus buffer size B (number of FDLs) is given in Fig. 22. It shows that the burst loss ratio is quite sensitive to the channel utilization/load. Moreover, the difference of burst loss ratios under random and self-similar traffic becomes more pronounced at low probability region, which seems consistent with Fig. 17. Note that the total buffer storage capacity does not increase linearly with B . The impact of number of data channels ($K - k$) per DCG on the burst loss ratio is depicted in Fig. 23. As expected, increasing the number of data channels per DCG can decrease the burst loss ratio. For high traffic load, the buffer size B should still be sufficiently large in order to achieve the targeted burst loss ratio. Since FDL optical buffer could be quite expensive, exploiting the multiplexing gain via channel grouping is very important in the design of optical routers and the OBS network.

C. BHP Delay and Offset-Time

The BHP delay is defined as the time elapsed from BHP entering the SCU (Fig. 8) until being processed by the BHP transmission (Tx) module. The BHP delay (variable part) is mainly determined by the traffic load, the forwarding time δ_f , the scheduler processing time δ_s , and the switch controller processing time δ_c . Knowing the distribution of the BHP delay is

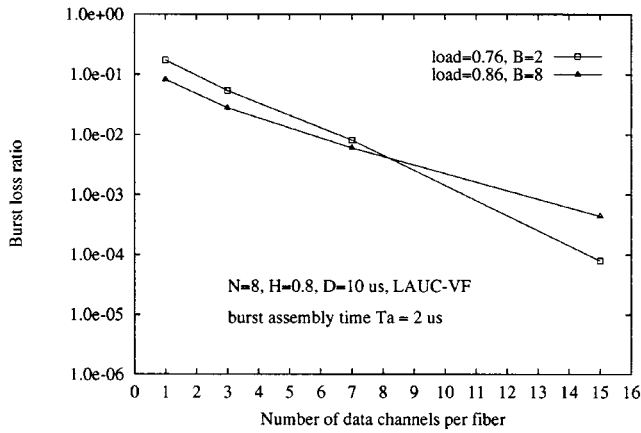


Fig. 23. Impact of number of data channels per DCG.

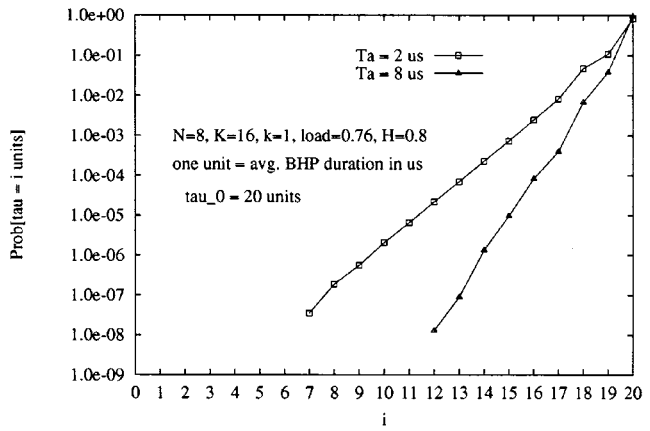


Fig. 25. The probability distribution of burst offset-time.

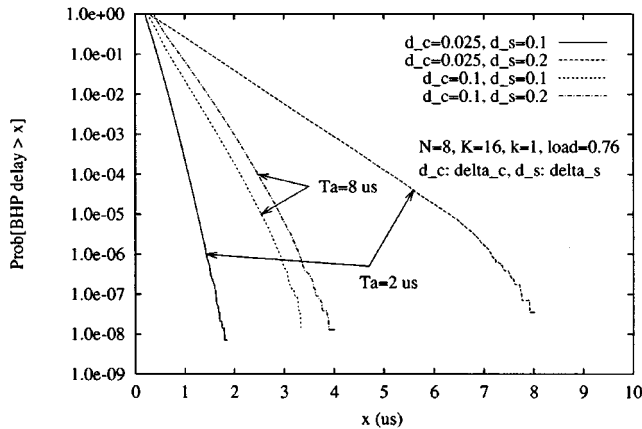


Fig. 24. The distribution of BHP delay.

critical to the dimensioning of the input FDL in Fig. 5. Under the same traffic condition as in Figs. 19 and 20, and given $\delta_f = 0.1 \mu\text{s}$, the impact of δ_s and δ_c on the BHP delay distribution is depicted in Fig. 24. For $T_a = 2 \mu\text{s}$ and $\delta_c = 0.025 \mu\text{s}$, the load of the scheduler is 0.52 when $\delta_s = 0.1 \mu\text{s}$ and increases to 0.87 when $\delta_s = 0.2 \mu\text{s}$. For $T_a = 8 \mu\text{s}$ and $\delta_c = 0.1 \mu\text{s}$, the load of the scheduler is 0.31 when $\delta_s = 0.1 \mu\text{s}$ and increases to 0.41 when $\delta_s = 0.2 \mu\text{s}$. Note that the BHP arrival rate to the SCU under $T_a = 2$ is almost 4 times as that under $T_a = 8$.

Use the average BHP duration in microseconds as a basic time unit, and set $\tau_0 = 20$ units. Fig. 25 shows an example of the distribution of burst offset-time τ . Because of the resynchronization, the majority of bursts still have the offset time $\tau = \tau_0$ after the first hop. It is clear from Fig. 25 that if we reduce the number of bursts in the network, which means longer burst assembly time T_a , there would be less BHP contention on outgoing CCGs, implying less variation in the burst offset-time. It is also found that the offset-time distribution will not change significantly after the second hop in the OBS network. The burst offset-time is an essential issue in the OBS. How to choose the initial burst offset-time τ_0 in an OBS network is not trivial and remains for further study.

VI. FURTHER DISCUSSIONS

Optical burst switching (OBS) provides an attractive alternative for realizing optical packet/burst switched WDM networks. In this paper, we present our initial work on the OBS network. There are still a lot of important issues that need to be solved before the OBS becomes practical. Burstification, burst offset-time management, data and control channel scheduling, and FDL buffer dimensioning are essential in OBS. Fast synchronization at the burst receiver of an edge router is also critical to the OBS concept and will certainly affect its efficiency. We found that these issues are closely related. For instance, FDL length depends on the burst length statistics, which are largely determined by the burstification. The burstification, the optical router size, the size of data and control channel groups will affect the burst offset-time management. The processing speeds of the scheduler and switch controller in the switch control unit (SCU) and the size of whole network will influence the burstification, etc. All these issues need to be carefully considered in the network design. Due to the limitations of FDL buffers, statistical multiplexing via channel grouping is vital in the optical burst-switched WDM network.

Viewing the OBS network as two coupled overlay networks—a pure optical network transferring data bursts, and a hybrid control network transferring burst headers (BHPs)—we find the coordination of these two networks to be crucial. Traffic and congestion control could be very challenging in the OBS networks since bursts can get lost due to the congestion in either the optical network or the control network or both. Ideally, we would like to design a control network that is congestion-free, or at least the loss of BHPs in the control network is kept extremely small. Network restoration, especially how to handle the failure of control channels, is also a critical issue.

How to support QoS and multicast traffic in the OBS are other issues that need to be studied. Due to the stringent time constraints of data bursts and their headers in optical routers plus the extremely high-speed environment, the roles of data/control channel scheduling algorithms could be quite limited compared to the scheduling algorithms in electronic routers. To support multicast traffic, both optical switch and the SCU should have multicast capability. To effectively coordinate them, new optical multicast switch architectures may be required.

REFERENCES

- [1] D. Blumenthal, P. Prucnal, and J. Sauer, "Photonic packet switches: Architectures and experimental implementation," *Proc. IEEE*, vol. 82, pp. 1650–1667, Nov. 1994.
- [2] S. Danielsen, B. Mikkelsen, C. Joergensen, and T. Durhuus, "WDM packet switch architectures and analysis of the influence of tuneable wavelength converters on the performance," *IEEE J. Lightwave Technol.*, vol. 15, pp. 219–227, Feb. 1997.
- [3] S. Danielsen, C. Joergensen, B. Mikkelsen, and K. Stubkjaer, "Optical packet switched network layer without optical buffers," *IEEE Photon. Technol. Lett.*, vol. 10, pp. 896–898, June 1998.
- [4] P. Gambini, *et al.*, "Transparent optical packet switching: Network architecture and demonstrators in the KEOPS project," *IEEE J. Select. Areas Commun.*, vol. 16, pp. 1245–1259, Sept. 1998.
- [5] L. Tamil, F. Masetti, T. McDermott, G. Castanon, A. Ge, and L. Tancevski, "Optical IP routers: Design and performance issues under self-similar traffic," *J. High Speed Networks*, vol. 8, pp. 59–67, Aug. 1999.
- [6] M. Yoo, M. Jeong, and C. Qiao, "A high speed protocol for bursty traffic in optical networks," in *Proc. SPIE'97 Conf. All-Optical Networking: Architecture, Control, Management Issues*, vol. 3230, Boston, Nov. 1997, pp. 79–90.
- [7] J. Turner, "Terabit burst switching," *J. High Speed Networks*, vol. 8, pp. 3–16, 1999.
- [8] C. Qiao and M. Yoo, "Optical burst switching (OBS)—A new paradigm for an optical internet," *J. High Speed Networks*, vol. 8, pp. 69–84, 1999.
- [9] F. Callegati, H. Cankaya, Y. Xiong, and M. Vandenhoute, "Design issues of optical IP routers for internet backbone applications," *IEEE Commun. Mag.*, pp. 124–128, Dec. 1999.
- [10] J. Kulzer and W. Montgomery, "Statistical switching architectures for future services," presented at ISS'84, Florence, Session 43A, May 7–11, 1984.
- [11] S. Amstutz, "Burst switching—An update," *IEEE Commun. Mag.*, pp. 50–57, Sept. 1989.
- [12] M. Yoo and C. Qiao, "A new optical burst switching protocol for supporting quality of service," in *Proc. SPIE'98 Conf. All Optical Commun. Syst.: Architecture, Control, Network Issues*, vol. 3531, Boston, Nov. 1998, pp. 396–405.
- [13] K. Claffy, G. Miller, and K. Thompson. The nature of the beast: Recent traffic measurements from an internet backbone. [Online]. Available: <http://ipn.nlanr.net/Papers/Inet98/index.html>
- [14] V. Paxson, "Fast approximation of self-similar network traffic," Tech. Rep. LBL-36750/UC-405, Apr. 1995.
- [15] <ftp://moat.nlanr.net/pub/MOAT/Traces>.
- [16] L. Tancevski, A. Ge, G. Castanon, and L. Tamil, "A new scheduling algorithm for asynchronous, variable length IP traffic incorporating void filling," in *Proc. OFC'99*.
- [17] Y. Xiong, M. Vandenhoute, and H. Cankaya, "Design and analysis of optical burst-switched networks," in *Proc. SPIE'99 Conf. All Optical Networking: Architecture, Control, Management Issues*, vol. 3843, Boston, MA, Sept. 19–22, 1999, pp. 112–119.
- [18] S. Q. Zheng, Y. Xiong, and M. Vandenhoute, "Hardware implementation of channel scheduling algorithms for optical routers with FDL buffers." To be submitted.
- [19] C. Metz, *IP Switching: Protocols and Architectures*. New York: McGraw-Hill, 1998.
- [20] J. Manchester, J. Anderson, B. Doshi, and S. Dravida, "IP over SONET," *IEEE Commun. Mag.*, pp. 136–142, May 1998.
- [21] M. Jeong, Y. Xiong, H. C. Cankaya, M. Vandenhoute, and C. Qiao, "Efficient multicast schemes for optical burst-switched WDM networks," in *Proc. IEEE ICC'00*, pp. 1289–1294.
- [22] P. Lin and R. Tench, "The exciting frontier of lightwave technology," *IEEE Commun. Mag.*, p. 119, Mar. 1999.



Yijun Xiong (M'93) received the B.S. and M.S. degrees from Shanghai Jiao Tong University, China, in 1984 and 1987, respectively, and the Ph.D. degree from the University of Ghent, Belgium, in 1994, all in electrical engineering.

Since 1998, he has been with the Alcatel Corporate Research Center in Richardson, TX. Previously, he was with the Research Center of Alcatel Bell Telephone (Antwerp, Belgium) from 1989 to 1992, the Laboratory for Communications Engineering, University of Ghent (Belgium) from 1992 to 1994, and INRS-Telecommunications (Montreal, Canada) from 1994 to 1996. From 1996 to 1998, he was an Assistant Professor of Electrical and Computer Engineering at Louisiana State University, Baton Rouge, LA. His current research interests include optical packet/burst networks, switch architectures, network survivability, traffic modeling, and network performance analysis.



Marc Vandenhoute (S'88–M'89) received the M.Sc. degree in computer engineering from Brussels Free University and the M.B.A. from Catholic University Leuven, both in Belgium.

He is in charge of the Network Architecture Unit of the Alcatel Corporate Research Center in Richardson, TX. His interests include optical internetworking active networks and QoS/multicast enhanced networks.



Hakki C. Cankaya (M'94) received the B.Sc. and M.Sc. degrees in computer engineering, both from Ege University, Izmir, Turkey, in 1990 and 1992, respectively, and the M.Sc. degree in computer science and the Ph.D. degree in computer engineering from Southern Methodist University, Dallas, TX, in 1995 and 1999, respectively.

He was a research assistant during his studies at Ege University and Southern Methodist University from 1990 to 1998. He is currently a Research Scientist at Network Architecture Department of AL-

CATEL Corporate Research Center in Richardson, TX. His recent research interests include all-optical IP based networks, teletraffic modeling and scheduling, and network survivability evaluation.

Dr. Cankaya is a member of IEEE Communications, Computer, and Reliability Societies, ACM, and Sigma Xi.