Optical Burst Switching (OBS) – A New Paradigm for an Optical Internet

Chunming Qiao[1] Myungsik Yoo

Dept of CSE Dept of EE

Lab for Advanced Network Design, Evaluation and Research (LANDER)
University at Buffalo
Buffalo, New York 14260

ABSTRACT

To support bursty traffic on the Internet (and especially WWW) efficiently, optical burst switching (OBS) is proposed as a way to streamline both protocol and hardware in building the future generation Optical Internet. By leveraging the attractive properties of optical communications and at the same time, taking into account its limitations, OBS combines the best of optical circuit-switching and packet/cell switching. In this paper, the general concept of OBS protocols and in particular, those based on Just-Enough-Time (JET), is described, along with the applicability of OBS protocols to IP over WDM. Specific issues such as the use of fiber delay-lines (FDL) for accommodating processing delay and/or resolving conflicts are also discussed. In addition, the performance of JET-based OBS protocols which use an offset time along with delayed reservation to achieve efficient utilization of both bandwidth and FDLs as well as to support priority-based routing is evaluated.

Keywords: bursty traffic, fiber delay line (FDL), IP, optical networks, reservation, WDM

[1] Contacting information: phone:(716)-645-3180 ext. 140, fax: (716)-645-3464 and email: qiao@computer.org. This research is sponsored in part by a grant from NSF under contract number ANIR-9801778

1 Introduction

The emergence of Terabit switches/routers, whose line speed has approached OC-48 (2.5 Gb/s) and may soon reach OC-192 (10Gb/s), makes it natural to provide direct WDM interconnects between these switches and routers, leading to current activities in building the so-called "Optical Internet" and "IP over SONET (over WDM)". These networks may be regarded as the first generation Optical Internet where switching is still performed in the electronic domain. In this paper, we study some issues related to using all-optical WDM networks as a layer directly beneath IP. Given the significant progress made and the continuing advances expected in the DWDM networking technology, which provides new and strong incentives to building a flexible, efficient and bandwidth-abundant fiber-optic network infrastructure capable of providing ubiquitous services, we believe that such future (2nd) generation Optical Internet will be not only desirable as a way to support a higher degree of data transparency for the benefit of certain applications but also *feasible* in the near future.

During the past several years, static and slowly reconfigurable WDM networks have been a focus, which is understandable given the constraints imposed by the devices and components, and the basic need to provide lightpaths to an upper layer such as SONET. In order for a WDM optical layer to provide differentiated services in an effective and feasible way, as well as to circumvent the current and/or fundamental economic and technological limits, we proposed an architectural framework allowing for several interoperable virtual optical networks (VONs) [27]. Under such a framework, each VON is allocated with appropriate resources (e.g., a subset of wavelengths) and applies either static or dynamic (i.e., adaptive) control, whichever is more appropriate. For example, a VON may adopt *On-demand reconfiguration* to support bursty traffic and short-lived connections (e.g, see [21,28]), while another VON may adopt *Self-reconfiguration* to support steady-traffic and long-lived connections (e.g., see [25, 29]). The former, called a dynamic VON, may use small but fast switches (e.g. based on either Lithium Niobate directional couplers or broadcast-and-select star couplers followed by SOAs) as well as wavelength converters. The latter, called a static VON, may use large but slow switches (e.g. opto-mechanical switches) without the wavelength conversion capability. It is expected that, as traffic nature changes from being voice-dominant to data-dominant, and at the same time, the device and component technology improves, the optical layer will evolve from perceived static VONs to a mixture of static and dynamic VONs.

In this paper, we will limit our attention to dynamic VONs, and in particular, those supporting bursty traffic. Such VONs can be referred to as "bursty" VONs which can be used for running

IP directly over WDM. There are two important drivers for building such an envisaged Optical Internet. One is the explosion of the data traffic over the Internet, especially the World-Wide-Web, which is bursty in nature; The other is the desire of the users/applications as well as the opportunities provided by the breakthrough made in the WDM optical networking technology to streamline both software (e.g. ATM signaling protocols) and hardware (e.g. SONET equipments) to reduce latency and cost.

The main contribution of this paper is the introduction of the new switching paradigm called optical burst switching (OBS). OBS can combine the best of the coarse-grained circuit-switching and the fine-grained packet-switching paradigms while avoiding their shortcomings, thereby efficiently supporting bursty traffic generated by upper level protocols or high-end user applications directly. Using OBS, a control (or set-up) packet is sent first, followed by a data burst on a separate wavelength. Such a one-way reservation paradigm is suitable for sending data requiring a high bit-rate and a low latency but having a relatively short duration compared to the end-to-end propagation delay of the network.

The rest of the paper is organized as follows. In the next section, we provide a general description of OBS protocols. In Sec. 3, we describe Just-Enough-Time (JET) [39,40], and also discuss several specific issues related to JET-based OBS protocols. In Section 4, we evaluate the performance of JET and its variations. We conclude the paper in Sec. 5.

2 Why OBS and What is OBS

In order to put optical burst switching (OBS) in perspective, we first describe a framework called *polymorphic control* of which OBS is an integral part.

2.1 Polymorphic Control

The framework of polymorphic control is a product of integrating many individual research ideas and results on optical network architectures, control and management. As mentioned earlier, under this framework, an optical layer is "sliced" into static and dynamic virtual optical networks (VONs), which apply *Self-reconfiguration* and *On-demand reconfiguration*, respectively.

One of the basic forms of Self-reconfiguration is scheduled communications [5, 9, 29, 34]. When the bandwidth (e.g. in terms of the number of wavelengths) in a static VON is limited, the set of communicating node pairs may be partitioned into a number of subsets such that the node pairs in each subset can communicate at the same time. Each subset is to be allocated a *super* time-slot during which data can be transmitted or received between the communicating nodes in that subset, and the number of time-slots determines the schedule length. A schedule specifying, among other things, the time-slot during which a given node pair can communicate, and the path and wavelength it will use, is then determined. Based on such a schedule, the VON can go through a pre-determined sequence of configurations by appropriately changing the switch settings inbetween two super time-slots. In this way, external electronic control and its associated implementation overhead and performance degradation are minimized.

In scheduled communications, two important performance measures are the schedule length and the bandwidth (i.e. wavelength) requirement, which relate to each other. With sufficient bandwidth, scheduled communications become *embedded* communications as a special instance, where the schedule length is one, or in other words, communications among the entire set of communicating nodes are accommodated at the same time [25, 29]. In a similar approach, which may also be considered as a form of Self-reconfiguration, a logical topology (analogous to a static VON) containing the set of communicating nodes is devised and embedded even when bandwidth is limited [22, 32], such that these nodes may communicate at the same time but a message from its source to its ultimate destination may go through more than one lightpaths, thus requiring O/E and E/O conversions at the nodes where two lightpaths meet.

In VONs adopting On-demand reconfiguration, where the performance measures include throughput, utilization, delay and blocking probability, dynamically changing traffic patterns are supported by transferring data in two basic fashions, namely circuit-switching and packet-switching. With circuit-switching, connections (or lightpaths) between source and destination pairs are established before data is transferred, and released after the transfer is completed. Both centralized control [1, 2, 7, 18] and distributed control [21, 26, 30, 42] have been studied, and in either case, it is common to use out-of-band signaling (i.e. a separate control network with a dedicated wavelength). With packet-switching [4, 8, 10], each intermediate node stores an incoming packet, and then forwards it to the next node based on its header and a locally stored routing table. Distributed control is natural and in-band signaling is more often used than out-of-band signaling. Note that alternately, a flow of packets can be switched based on the match between a label carried by each packet's header and a label stored at each node, which is set up either by previous packets of the same flow (as in IP-switching [23]) or by the network (as in Tag-switching [33]). A bursty VON is a dynamic VON that adopts a novel paradigm called optical burst switching (OBS), which can be used to support MPLS (Multi-protocol Label Switching) in an IP over WDM environment.

2.2 Motivation

The main motivation for considering *optical burst switching* (OBS) is that some traffic in broadband multimedia services is inherently bursty. More specifically, recent studies have shown that, in addition to traffic in a local Ethernet and between remote Ethernets (i.e. WAN traffic), traffic generated by Web browsers, wide-area TCP connections (including FTP and TELNET traffic carried over TCP connections), and variable-bit-rate (VBR) video sources are all *self-similar* (or bursty at all time scales) [3, 19]. More importantly, some studies have concluded that, contrary to the common assumption based on Poisson traffic, multiplexing a large number of self-similar traffic streams results in bursty traffic [13, 24].

Existing switching paradigms in optical networks are not suitable for supporting bursty traffic. Specifically, using optical circuit-switching via wavelength routing [7, 22, 31], a lightpath needs to be established first from a source node to a destination node using a dedicated wavelength on each link along a physical path. The bandwidth, therefore, would not be efficiently utilized if the subsequent data transmission does not have a long duration relative to the set-up time of the lightpath. In addition, given that number of wavelengths available is limited, not every node can have a dedicated lightpath to every other node, and accordingly, some data may take a longer route and/or go through O/E and E/O conversions. Furthermore, the extremely high degree of transparency of the lightpaths limits the network management capabilities (e.g. monitoring and fast fault recovery).

An alternative to optical circuit switching is optical or photonic packet/cell switching in which a packet is sent along with its header [4,8,10]. While the header is being processed by an intermediate node, either all-optically or electronically (after an O/E conversion), the packet is buffered at the node in the optical domain. However, high-speed optical logic, optical memory technology, and synchronization requirements are major problems with this approach. In particular, the limited buffering time that can be provided to optical signals prevents worm-hole routing and virtual cut-through routing [11, 17], which are popular in systems with electronic buffers, from being deployed effectively in optical networks.

In order to provide high-bandwidth transport services at the optical layer for bursty traffic in a flexible, efficient as well as feasible way, what is needed then is a new switching paradigm that can leverage the attractive properties of optical communications, and at the same time, take into account its limitations. *Optical burst switching* (OBS) is intended to accomplish exactly that.

2.3 An Overview of OBS

In OBS, a control packet is sent first to set up a connection (by reserving an appropriate amount of bandwidth and configuring the switches along a path), followed by a burst of data without waiting for an acknowledgement for the connection establishment. In other words, OBS uses one-way reservation protocols similar to tell-n-go (TAG) [37,38], also known in ATM as *fast reservation protocol* (FRP) or ATM Block Transfer with Immediate Transmissions (or ABT-IT) [16,35]. This distinguishes OBS from circuit-switching as well as from other burst-switching approaches using protocols such as Reservation/scheduling with Just-In-Time switching (RIT) [15] and tell-and-wait (TAW), also known in ATM as ABT-DT (Delayed Transmissions) [6,36], all of which are two-way reservation protocols.

OBS also differs from optical or photonic packet/cell switching mainly in that the former can switch a burst whose length can range from one to several packets to a (short) session using one control packet, thus resulting in a lower control overhead per data unit. In addition, OBS uses out-of-band signaling, but more importantly, the control packet and the data burst are more loosely coupled (in time) than in packet/cell switching. In fact, they may be separated at the source as well as subsequent intermediate nodes by an offset time as in the Just-Enough-Time (JET) protocol to be described later. By choosing the offset time at the source to be larger than the total processing time of the control packet along the path [39,40], one can eliminate the need for a data burst to be buffered at any subsequent intermediate node just to wait for the control packet to get processed. Alternatively, an OBS protocol may choose not to use any offset time at the source, but instead, require that the data burst go through, at each intermediate node, a *fixed* delay that is no shorter than the maximal time needed to process a control packet at the intermediate node. Such OBS protocols will be collectively referred to as TAG-based since their basic concepts are the same as that of TAG itself.

One way to support IP over WDM using OBS is to run IP software, along with other control software as a part of the interface between the network layer and the WDM layer, on top of every optical (WDM) switch. In the WDM layer, a dedicated control wavelength is used to provide the "static/physical" links between these IP entities. Specifically, it is used to support packet switching between (physically) adjacent IP entities which maintain topology and routing tables. To send data, a control packet is routed from a source to its destination based on the IP addresses it carries (or just a lable if MPLS is supported) to set up a connection by configuring all optical switches along the path. Then, a burst (e.g. one or more data IP packets, or an entire message) is delivered without going through intermediate IP entities, thus reducing its latency as well as the processing load at the IP layer. Note that, due to the limited "opaqueness" of the control packet, OBS can achieve a high degree of adaptivity to congestions or faults (e.g,. by using deflection-routing), and support priority-based routing as in optical cell/packet switching, as to be discussed later.

In OBS, the wavelength on a link used by the burst will be released as soon as the burst passes through the link, either automatically according to the reservation made (as in JET) or by an explicit release packet. In this way, bursts from different sources to different destinations can effectively utilize the bandwidth of the same wavelength on a link in a time-shared, statistical multiplexed fashion. Note that, in case the control packet fails to reserve the bandwidth at an intermediate node, the burst (which is considered *blocked* at this time) may have to be dropped. OBS can support either reliable or unreliable burst transmissions at the optical layer. In the former, a negative acknowledgement is sent back to the source node, which retransmits the control packet and the burst later. Such a retransmission may be necessary when OBS is to support some application protocols directly, but not when using an upper layer protocol such as TCP which eventually retransmits lost data.

In either case, a dropped burst wastes the bandwidth on the partially established path. However, since such bandwidth has been reserved exclusively for the burst, it would be wasted even if one does not send out the burst (as in two-way reservation). Similar arguments apply to optical or photonic packet switching as well. In order to eliminate the possibility of such bandwidth waste, a blocked burst (or an optical packet) will have to be stored in an electronic buffer after going through O/E conversions, and later (after going through E/O conversions), relayed to its destination. Fiber-optical delay lines (FDLs) providing limited delays at intermediate nodes, which are not mandatory in OBS when using the JET protocol, would help reduce the bandwidth waste and improve performance in OBS as to be discussed next. Note that, when using TAG-based OBS protocols (or optical/photonic packet switching), FDLs (or optical buffers) are required to delay each optical burst when the control packet (or the packet header) is processed, but do not help improve performance.

Summarizing the above discussions, switching optical bursts achieves, to certain extent, a balance between switching coarse-grained optical circuits and switching fine-grained optical packets/cells, and combines the best of both paradigms, as illustrated in Table 1.

Optical Switching (paradigm)	Bandwidth Utilization	Latency (set-up)	Optical Buffer	Proc./Sync. Overhead (per unit data)	Adaptivity (traffic & fault)
Circuit	low	high	not required	low	low
Packet/Cell	high	low	required	high	high
Burst	high	low	not required	low	high

Table 1: A comparison between three optical switching paradigms

3 The JET Protocol And Its Variations

The proposed *Just-Enough-Time* (or JET) protocol for OBS has two unique features, namely, the use of delayed reservation (DR) and the capability of integrating DR with the use of FDL-based buffered burst multiplexers (BBMs), which are to be described in this section. These features make JET and JET-based variations especially suitable for OBS when compared to TAG-based OBS protocols and other one-way reservation based OBS protocols that lack either or both features.

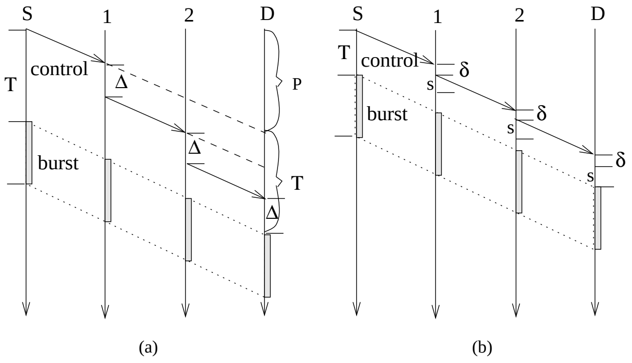

Figure 1: OBS using the JET protocol

Figure 1 illustrates the basic concept of JET. As shown, a source node having a burst to transmit first sends a control packet on a signaling channel (which is a dedicated wavelength) towards the destination node. The control packet is processed at each subsequent node in order to establish an all-optical data path for the following burst. More specifically, based on the information carried in the control packet, each node chooses an appropriate wavelength on the outgoing link, reserves the bandwidth on it, and sets up the optical switch. Meanwhile, the burst waits at the source in the *electronic* domain. After an *offset time*, T, whose value is to be determined next, the burst is sent in optical signals on the chosen wavelength (at say, 2.5Gb/s).

3.1 The Use of Offset Time

For simplicity, let us assume that the time to process the control packet, reserve appropriate bandwidth and set up the switch is Δ time units at each node, and ignore the receiving and transmission time of the control packet. In a TAG-based OBS protocol (or optical or photonic packet switching), a burst is sent by the source along with the control packet without any offset time (i.e., $T = 0$ in Figure 1). In addition, at each subsequent intermediate node, the burst waits for the control packet to be processed, and the two are sent to the next node without any offset time either. In this way, both the control packet and the burst will be delayed for Δ units, which will be referred to as the *per-node control latency.* Accordingly, the minimum latency of the burst including the total propagation time, denoted by P, but excluding its transmission time, is $P + \Delta \cdot H$, where H is the number of hops along the path (e.g., in Figure 1, $H = 3$).

In JET, we can choose the offset time T to be $\Delta \cdot H$, as shown in Figure 1 (a), to ensure that there is enough headroom for each node to complete the processing of the control packet before the burst arrives. In this way, the burst will not encounter a longer latency than using TAG-based OBS protocols. In fact, we may partition Δ into roughly two parts: let δ be the time to process the control packet and *initiate* other operations such as switch setting, and $s = (\Delta - \delta)$ be the time required to finish these operations. Using JET, the control packet can be sent out to the next node immediately after spending δ time units at each intermediate node, or in other words, s time units before the burst. This effectively overlaps the switching setting time at a node with the time for the control packet to propagate to (and possibly get processed at) the subsequent node. Consequently, one can reduce the offset time to $T' = \delta \cdot H + s$, and latency to $P + T'$, which are $s \cdot (H - 1)$ smaller than those using TAG-based OBS protocols. In the rest of the paper, we will ignore the difference between Δ and δ (and between T and T'), and just use "processing delay" to refer to the per-node control latency.

It is important to note that the burst can be sent without having to wait for an acknowledgement from its destination. At 2.5 Gb/s, a burst of 500 Kbytes (or 4,000 average-sized IP packets) can be transmitted in about $1.6ms$. However, an acknowledgement would take $2.5ms$ just to propagate over a distance of merely 500km. This explains why one-way reservation protocols are generally better than their two-way counterparts for bursty traffic over a relatively long distance. Once a burst is sent, it *passes through* the intermediate nodes without going through any buffer, so the minimal latency it encounters would be the same as if the burst is sent along with the control packet as in optical packet switching. Of course, if a burst is extremely small, one may just as well send the data along with the control information using packet-switching.

3.2 DR for Efficient Bandwidth Utilization

Figure 2 (a) illustrates why delayed reservation (DR) of bandwidth is useful in achieving efficient bandwidth utilization. Using a TAG-based OBS protocol, the bandwidth on the outgoing link is reserved from t'_1, the time node X finishes the processing of the (first) control packet. In JET, one may also reserve the bandwidth in the same way. However, it is natural to *delay* the bandwidth reservation till t_1, the time the (first) burst arrives. Here, $t_1 > t'_1$ and their difference is the value of the offset time between the burst and its corresponding control packet at node X.

Note that, a way to determine the arrival time of a burst, e.g. t_1, when the processing time of a control packet may vary from one node to another, is to let the control packet carry the value of the offset time to be used at the next node. This value can be updated based on the processing time encounted by the control packet at the current node. In the above example, immediately after the control packet succeeds in reserving the bandwidth, its transmission is scheduled, say, at t''_1. The value of the offset time to be used at the next node is then obtained by subtracting $t''_1 - t'_1$ from the current value. Obviously, some guardbands around the bursts may still be needed to accommodate possible jitters but due to the limited space, such a topic will not be addressed in this paper.

In addition to taking into account the arriving time of the burst, t_1, what is more important is that in JET, the bandwidth may be reserved until $t_1 + l_1$, where l_1 is the burst duration, instead of until *infinity*. This will increase the bandwidth utilization and reduce the probability of having to drop a burst. For example, in both cases shown in Figure 2(a), namely $t_2 > t_1 + l_1$ and $t_2 < t_1$, respectively, the second burst will be dropped at node X if has no buffer for the burst when using

TAG. However, when using JET, the second burst will not be dropped in case 1, nor in case 2, provided that its length is shorter than $t_1 - t_2$.

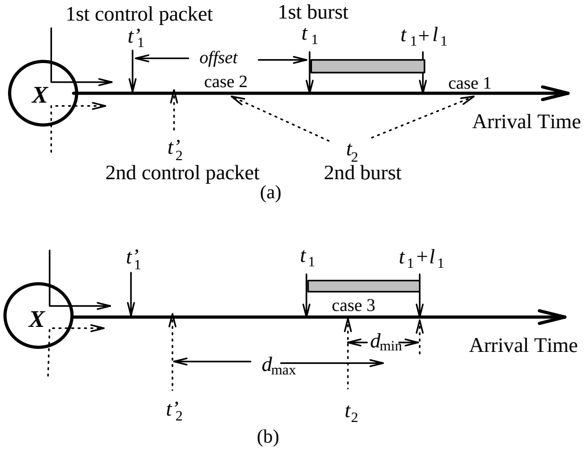

Figure 2: Delayed reservation (DR) and its usefulness with or without buffer

Note that, DR goes hand in hand with the use of offset time. In addition, although burst length may vary, we may assume that the length of a burst is known before the corresponding control packet is sent. This assumption is natural in some applications such as file transfer or WWW downloading. However, if the burst length is unknown, one may delay the control packet until either the entire burst arrives (from an upper layer), or a certain length is reached. To take advantage of the use of an offset time in JET, thereby reducing the pre-transmission latency, an alternative is to send out the control packet as soon as possible by using an estimated value of the burst length. If it is an over-estimation, another control (release) packet may be sent to release the extra bandwidth reserved. If it is an under-estimation, then the remaining data will be sent as one or more additional bursts. JET may also support an entire session by reserving the bandwidth to infinity, and use an explicit release packet when the circuit is no longer needed (i.e. the session ends).

3.3 Intelligent Buffer Management

As mentioned earlier, JET does not mandate the use of buffer, nor its size in terms of the maximal number of bursts or bits that can be stored (or delayed) simultaneously, and/or maximal delay it can provide to each burst at a node. Nevertheless, the dropping probability can be further reduced, and both bandwidth utilization and performance can be further improved, if a burst can be *buffered* (or delayed) at an intermediate node. Figure 3 shows two possible designs of buffered burst multiplexer (BBM) based on fiber-delay lines (FDLs), which can be used in an "output-buffered" photonic switch/router. In either case, a burst coming from any of the three inputs can be buffered for a maximal of $B = 2^{n+1} - 1$ time units. The difference between the shared BBM shown in Figure 3 (a) and the dedicated BBM shown in Figure 3(b) is that the latter is more complex (and costly) but more powerful. Again, due to the limited space, we will not address the issue of designing cost-effective BBMs further in this paper. Nevertheless, we note that bursts at different wavelengths may share the same FDL, and adding wavelength converters in a node may increase the utilization of the FDLs even further [12,14]. With current technology, a FDL-based buffer containing a few kilometers of fiber, and thus providing a few tens of μs delay is feasible [20].

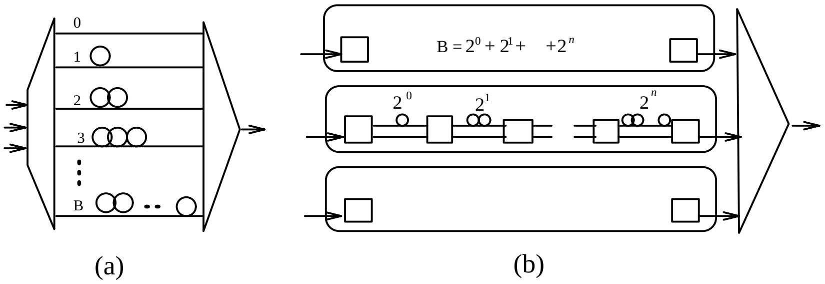

Figure 3: An example of (a). a shared BBM and (b). a dedicated BBM

Given that optical buffer is a scarce and expensive resource, JET makes effective use of buffer in two ways. Firstly, in JET, each burst will wait (i.e. be buffered) at the source in the electronic domain during the offset time and there is no need to buffer it at any intermediate node at all when its control packet encounters no blocking along the path. Secondly, in case a control packet is blocked at an intermediate node, DR and BBM can be integrated in that DR is also useful in increasing the effectiveness of the available buffer through intelligent buffer allocation and management, just as DR is useful in increasing the bandwidth utilization as discussed in the previous subsection. For example, refer to case 3 shown in Figure 2 (b) where $t_1 < t_2 < t_1 + l_1$. Clearly, if the second burst can be delayed (or buffered) for at least $d_{min} = t_1 + l_1 - t_2$, then it needs not to be dropped. In JET, node X *can* determine if it has a sufficient number of FDLs in the BBM at the output port. If so, it reserves a minimum number of FDLs necessary to provide a delay $d \geq d_{min}$ using DR (that is, reserve the FDLs only for the period from t_2 to $t_2 + l_2$ instead of from t'_2 to infinity); Otherwise, no FDL will be reserved. Note that, without using DR, the node cannot know how much delay is actually necessary. Consequently, it will have to reserve the entire buffering capacity available in the BBM at t'_2, say, d_{max} ($\leq B$) as depicted in Figure 2 (b). Such a *blind and brutal-force* reservation will waste some of the buffering capacity when d_{max} turns out to be excessive (i.e. if $t'_2 + d_{max} > t_1 + l_1$). Even worse, it will waste the entire buffering capacity when d_{max} turns out to be insufficient (i.e. if $t'_2 + d_{max} < t_1 + l_1$) since the burst will have to be dropped after all.

3.4 Adaptive Routing and Priority Schemes

A critical design issue in OBS is how to reduce burst dropping probability. With no or very limited buffering, the dropping probability of a burst may be improved by implementing adaptive routing and/or assigning it with a higher priority.

As mentioned earlier, a TAG-based OBS protocol does not use any offset time. Instead, a data burst goes through a fixed delay (FDL) at each intermediate node to account for the processing delay encounted by the corresponding control packet. This facilitates the use of a different path to a given destination each time a source sents a new burst (or retransmits a dropped burst), as well as deflection-routing at intermediate nodes when a burst is blocked.

A JET-based OBS protocol can also support multi-path routing from a given source to a given destination as long as the (approximate) number of hops along each path is known. To support deflection-routing at an intermediate node when there is no bandwidth to reserve on the primary outgoing link, the control packet chooses an alternate outgoing link, and sets the switch accordingly so that the data burst will also follow the alternate path. If a minimal offset time based on the primary path was used, and the alternate path is longer (in terms of number of hops), then the data burst needs to be delayed further in order to make up for the increase in the total processing delay encounted by the control packet along the alternate path. This can be accomplished by letting the data burst go through some FDLs at one or more nodes before the offset time goes to zero, even if no blocking occurs at these nodes. We note that a JET-based protocol can support limited adaptivity even without using FDLs. Specifically, one can use an extra offset time at the source to account for a possible increase in the total processing delay of the control packet due to deflection routing.

In addition to being useful for deflection-routing, having an additional offset time can increase the priority of a burst. This is because the corresponding control packet will likely to succeed in reserving the bandwidth into the future, given that very few other control packets arriving earlier (or around the same time) might have reserved (or want to reserve) the bandwidth that much in advance. This property of an additional offset time can be utilized to improve *fairness* by assigning a higher priority to bursts which must travel for a longer distance (in terms of the number of hops) from their sources to destinations. We will call this variation of JET which implements such a priority scheme JET-FA (for fairness), whose performance will be shown in the next section.

4 Performance Evaluation

To evaluate the performance of JET and its variations, we assume that bursts have an exponentially distributed duration with an average denoted by L, and that the destination of each burst is evenly distributed among all nodes except its source. For a burst whose source-destination distance (or path length) is H (*hops*), the minimum offset time is set to $\Delta \cdot H$. We will use k to denote the number of channels (e.g. wavelengths) per link, b to denote the ratio of the maximal buffering time (i.e B in Figure 3) to L, and c to denote the ratio of Δ to L.

4.1 The Effectiveness of DR

We first report simulation results that compare the performance of the JET protocol (which uses DR) with a variation called NoDR, which does not uses DR, although both use the same (minimum) offset time.

A torus (or meshed-ring) network having $N = n^2$ nodes which may be considered as having n horizontal rings and n vertical rings is simulated. Each node in the network is connected to a local host and four other nodes, and the propagation delay between any two nodes is denoted by p. To send a burst or control packet, a shortest path from its source to its destination will be used, whose average path length, H_{avg}, is approximately $n/2$. When multiple shortest paths exist in a torus, one is randomly chosen for each burst. In addition, loss-less communications is assumed in which a dropped burst is retransmitted after a random *backoff* time, whose value is evenly distributed between 0 and $2L$ with an average of L.

The performance metrics used are (average) link utilization, denoted by U, as a function of the (average) delay of a burst due to blocking, denoted by D. Note that, based on such a definition, the delay, D, will be 0 if no blocking occurs using an one-way reservation protocol and a minimum offset time value. As a comparison, using a two-way reservation protocol, the minimum delay will be about $H_{avg} \cdot (p + \Delta)$, which is the time for an acknowledgment to be received by the source.

Simulation results have been obtained using the following *default* parameter values: $N = 4 \times 4$,

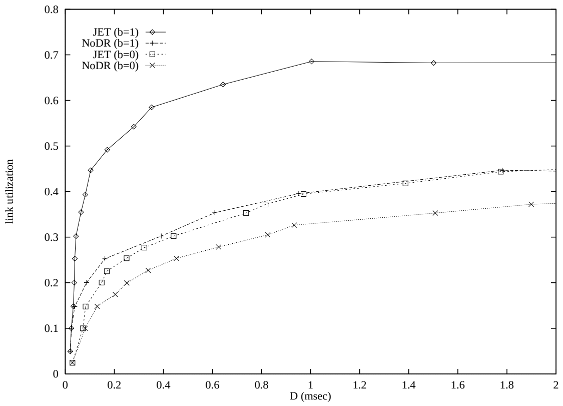

Figure 4: Performance evaluation of various protocols

$k = 4$, $c = 0.1$, and $b = 1$ if BBMs are used (otherwise, $b = 0$). In addition, we let $L = 40\mu sec$, which is equivalent to a burst of $100K$bits (or 100 average-sized IP packets) transmitted at 2.5Gb/s, and $p = 2.5msec$, which is equivalent to a link length of 500km. The link utilizations achieved by JET and NoDR with or without using BBMs are shown in Figure 4. Note that, if different L and p are used, the shape and the relative position of the four curves in the Figure will not change much, only the scale of the X-axis (the delay D) will.

The results indicate that JET which uses DR alone (i.e. $b = 0$) can achieve the same performance as NoDR which uses BBMs alone (i.e. $b = 1$). In addition, JET ($b = 1$) can outperform NoDR ($b = 0$) by about 80%, and the other two by at least 50%. Since the performance improvement of JET ($b = 1$) over NoDR ($b = 0$) is larger than the sum of the improvement of NoDR ($b = 1$) over NoDR($b = 0$) and the improvement of JET($b = 0$) over NoDR ($b = 0$), one may conclude that the use of DR can improves not only the bandwidth utilization, but also the buffer effectiveness (through intelligent buffer allocation and management).

We have also compared JET with TAG-based OBS protocols, and the results (although not shown) have indicated that there is no significant difference between the performance of the two as long as c is small (in fact, their performance will be the same when $c = 0$). If, however, c is large (e.g. $c \approx 1$), and the FDLs, which are required by the TAG-based OBS protocols to delay a burst while its corresponding control packet is being processed, can be used in JET for resolving conflicts, JET may outperform TAG-based protocols simply because the former will have an effectively larger b.

4.2 Scalability Analysis

In this subsection, we compare JET with NoDR by varying the values of the parameters k, b, c and N. Note that since both JET and NoDR will use the same b value ($b = 1$ by default), any performance improvement of JET over NoDR will be entirely due to the use of DR.

Figure 5 shows the link utilization improvement *ratio* of JET over NoDR as k varies. The absolute values of link utilization U achieved by the two protocols under various k are also shown at right as a reference. These results have been obtained under the assumption that when $k > 1$, wavelength conversions can be performed so that a burst can go out on any wavelength that is free on the outgoing link. This is why U increases with k in both protocols. However, one may also observe that in JET, such an increase in utilization is larger than in NoDR, and as a result, the improvement ratio is higher when $k > 1$ than when $k = 1$. This can be explained as follows. When all the wavelengths on the outgoing link are in use, and thus a blocking occurs, a control packet in JET can pick a wavelength on which the bandwidth will be released first based on the known release time of each wavelength, and reserve an appropriate amount of buffer. On the other hand, in NoDR, the blocked control packet cannot predict which wavelength will be available first, and thus can take little advantage of the multiple choices among wavelengths.

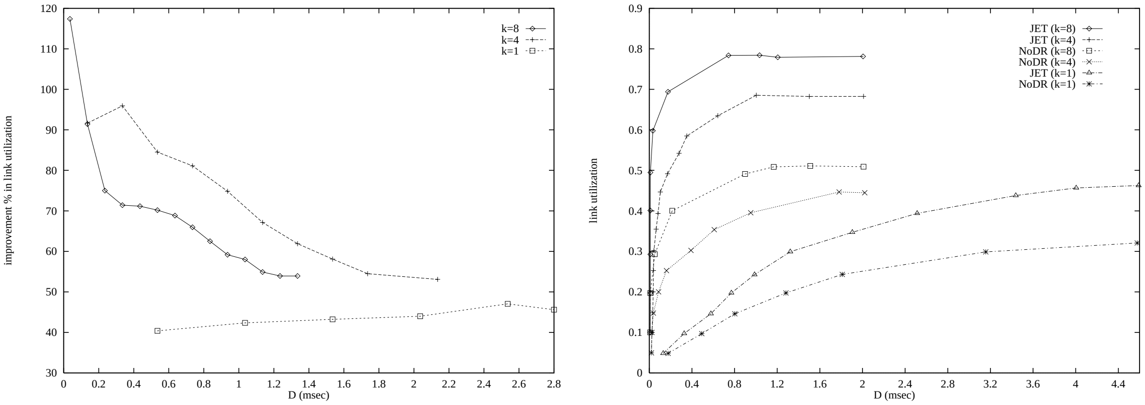

Figure 5: Bandwidth utilization improvement of JET over NoDR when $k = 1$, 4 and 8.

From Figure 5, one may also observe that when $k = 4$, the improvement ratio initially increases to close to 100% with the delay, then gradually decreases, and finally settles down at 50%. This can be explained by examining the right half of the figure. Specifically, this is caused by the following "push-and-pull" effects. One is that U increases with D in NoDR (albeit in JET as well), and the other is that the rate at which U increases slows down as D increases beyond a certain value. For similar reasons, the improvement ratio when $k = 8$ is lower than when $k = 4$ for most values of D.

The utilization improvement of JET over NoDR when b varies is plotted in Figure 6. As can be seen, the improvement ratio is higher for a larger b value, but converges to around 50% for all b values when traffic is heavy. Note that as with a larger k, a larger b leads to a higher absolute value of U, which suggests that JET is scalable to both k and the maximal buffering time.

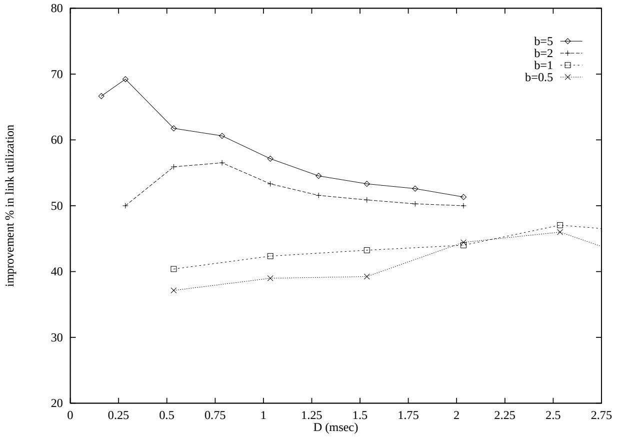

Figure 6: Bandwidth utilization improvement ratio when $b = 0.5$, 1, 2, and 5.

Figure 7 shows the utilization improvement ratio of JET over NoDR for different values of N. The peak improvement ratio is higher for a larger N, and this is because as N increases, so is H_{avg}, which results in a longer offset time used in JET, and thus a larger amount of bandwidth (and buffering capacity) saved when compared to NoDR. Note that the peak improvement ratios of 44%, 55% and 75% are obtained at $D = 2.5msec$, $5msec$ and $7.5msec$, respectively, in the 4×4, 8×8 and 12×12 tori. These values of D are still low when compared to the minimum delay of a two-way reservation protocol (which would be about $5msec$, $10msec$ and $15msec$, respectively, in these tori).

The effect of c on the utilization improvement ratio of JET protocol over NoDR is shown in Figure 8 where c varies from 0.1 to 1. A larger c means a relatively larger Δ, and thus a larger amount of offset time. It also means that more bandwidth will be wasted in NoDR (as discussed in the case for a larger N). Hence, as expected, when c increases, the improvement ratio increases proportionally. In particular, when $c = 1$, the improvement ratio can reach up to 400%. Since the transmission speed may increase much faster than the processing speed as in the past few decades, one may expect c will increase with time, making JET more attractive and desirable.

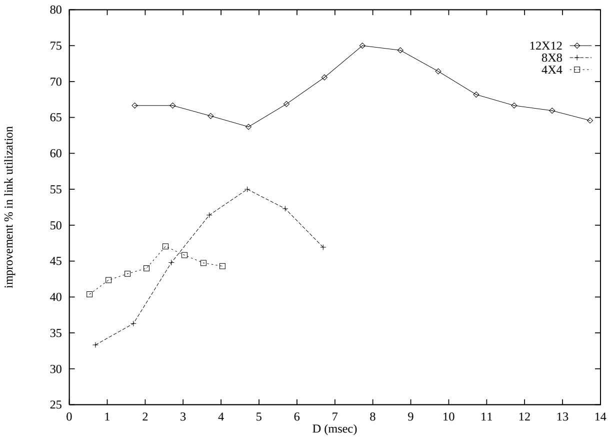

Figure 7: Bandwidth utilization improvement ratio when $N = 4 \times 4$, 8×8 and 12×12.

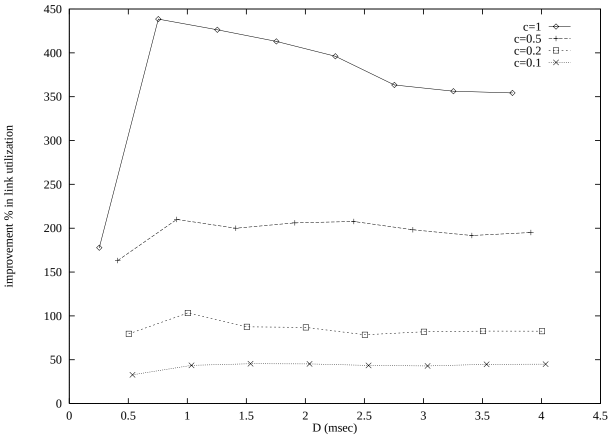

Figure 8: Bandwidth utilization improvement ratio when $c = 0.1$, 0.2, 0.5, and 1

4.3 QoS Support

In this subsection, we show how an additional offset time can be used to support priority and support QoS (e.g., to achieve fairness) without requiring the use of buffer at intermediate nodes. The performance of the JEt-FA protocol is evaluated under a slightly different model in which a blocked burst will be dropped and not retransmitted. Figure 9 shows the average dropping probability of a burst as a function of the distance between its source and destination, denoted by H, when $k = 8$, $b = c = 0$, and the traffic load (relative to link capacity) is 0.5 and 0.6, respectively. Two basic JET-FA protocols which assign an additional offset time which is a multiple of H (and L) are simulated. Specifically, in JET-FA(5L) and JET-FA(10L), a burst traveling H hops uses an additional offset time of $5L \times H$ and $10L \times H$, respectively. Note that, when $L = 40\mu sec$ and

$H = 6$, the maximal extra offset time would be only $6 \times 10L$ or 2.4 $msec$, which is quite tolerable for bursts that have to travel for 6 hops with an average per-hop distance of, say 500 km. As can be seen, the fairness is improved over the JET protocol. However, as a trade-off, the throughput of JET-FA can be slightly lower than JET. Nevertheless, with a large enough k, JET-FA can achieve approximately the same throughput as JET.

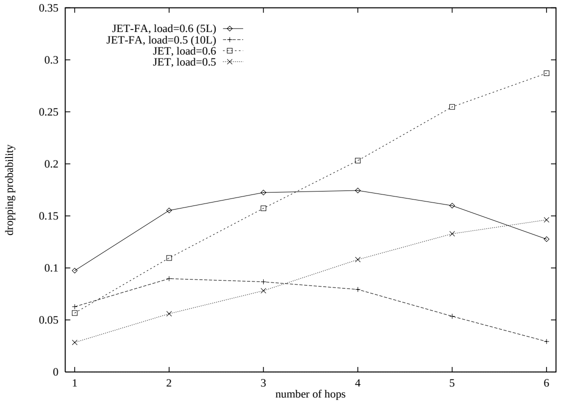

Figure 9: Improve the fairness using JET-FA protocols

Note that, we may apply the same idea in order to provide differentiated services without using buffer in a WDM network. Specifically, we can use an additional offset time when sending every burst belonging to a high priority class. For example, assume that the durations of the low priority bursts have an exponential distribution with an average L. If the additional offset time used is $5L$, which is longer than at least 99% of the low-priority bursts, then at most 1% of the low-priority bursts may block a high priority burst. As a result [41], the average dropping probability of the high priority bursts will be at least 10 times lower when $k = 8$, and several orders of magnitude lower when $k = 32$ (although the average blocking probability of all bursts remains unchanged).

5 Concluding Remarks

In this paper, we have proposed a novel paradigm called optical burst switching (OBS) as an efficient way to support the bursty data traffic, e.g. IP traffic, on top of WDM networks. We have described two major types of OBS protocols, both of which use an out-of-band control packet to set up the optical switches for the following data burst. One, which is based on tell-and-go (TAG), delays a burst at every intermediate node using, for example, fiber-delay lines (FDLs). The other, which is based on Just-Enough-Time (JET), delays the transmission of the burst at its source by an offset time, and thus can either eliminate the need for or make more efficient use of FDLs. We have also proposed a time-stamping technique to facilitate the use of an offset time and delayed reservation in JET-based protocols.

Performance evaluation results have indicated that JET-based OBS protocols can achieve a good bandwidth utilization by using delayed reservation, and improve fairness by assigning an additional offset time (which is equivalent to a higher priority) to bursts traveling through more hops. As a future direction, we note that, with the limited degree of opaqueness provided by control packets, and the ability to achieve better utilization of the network resources, OBS can be used to efficiently support multicasting at the optical layer to take advantage of the inherent multicasting capability of some optical switches as well as the knowledge of the physical topology of the WDM layer.

Acknowledgement

The authors would like to thank Yousong Mei and Xijun Zhang of LANDER at University of Buffalo for their valuable comments.

References

[1] K. Bala, T. Stern, and K. Bala. Algorithms for routing in a linear lightwave network. In *Proceedings of the IEEE InfoCom*, pages 1–9, 1991.

[2] R. A. Barry and P. A. Humblet. Models of blocking probability in all-optical networks with and without wavelength changers. In *Proceedings of IEEE Infocom*, pages 402–412, April 1995.

[3] J. Beran, R. Sherman, M. Taqqu, and W. Willinger. Long-range dependence in variable-bit-rate video tarffic. *IEEE Transactions on Communications*, 43:1566–1579, 1995.

[4] D. Blumenthal, P. Prucnal, and J. Sauer. Photonic packet switches - architectures and experimental implementations. *IEEE Proceedings*, 82:1650–1667, November 1994.

[5] M.S. Borella and B. Mukherjee. Efficient scheduling of nonuniform packet traffic in a WDM/TDM local lightwave network with arbitrary transceiver tuning latencies. *IEEE Journal on Selected Areas in Communications*, 14:923–934, 1996.

[6] P. E. Boyer and D. P. Tranchier. A reservation principle with applications to the ATM traffic control. *Computer Networks and ISDN Systems*, 24:321–334, 1992.

[7] I. Chlamtac, A. Ganz, and G. Karmi. Lightpath communications: an approach to high-bandwidth optical WANs. *IEEE Transactions on Communications*, 40:1171–1182, July 1992.

[8] I. Chlamtac et al. Cord: contention resolution by delay lines. *IEEE Journal on Selected Areas in Communications*, 14:1014–1029, June 1996.

[9] H.S. Choi, H.-A. Choi, and M. Azizoglu. Efficient scheduling of transmission in optical broadcast networks. *IEEE/ACM Transactions on Networking*, 4:913–920, 1996.

[10] R. Cruz and J.-T Tsai. Cod: alternative architectures for high speed packet switching. *IEEE/ACM Transactions on Networking*, 4:11–21, February 1996.

[11] W.J. Dally and C.L. Seitz. Deadlock-free message routing in multiprocessor interconnection networks. *IEEE Transactions on Computers*, 36:547–553, May 1987.

[12] S. Danielsen et al. WDM packet swicth architectures and analysis of the influence of tunable wavelength converters on the performance. *IEEE/OSA Journal of Lightwave Technology*, 15:219–227, February 1997.

[13] A. Erramilli, O. Narayan, and W. Willinger. Experiemental queueing analysis with long-range dependent packet traffic. *IEEE/ACM Transactions on Networking*, 4(2):209–223, 1996.

[14] J. Gabriagues and J. Jacob. Exploitation of the wavelength domain for photonic switching in IBCN. In *Proc. of ECOC*, volume 2, pages 59–66, 1991. Paris, France.

[15] G.C. Hudek and D.J. Muder. Signaling analysis for a multi-switch all-optical network. In *Proceedings of Int'l Conf. on Communication (ICC)*, pages 1206–1210, June 1995.

[16] ITU-T Rec. I.371. Traffic control and congestion control in B-ISDN. Perth, U.K. Nov. 6-14, 1995.

[17] P. Kermani and L. Kleinrock. Virtual cut-through : A new computer communication switching technique. *Computer Networks*, 3:267–286, 1979.

[18] K.C. Lee and Victor O.K. Li. A circuit rerouting algorithm for all-optical wide-area networks. In *Proceedings of IEEE Infocom*, pages 954–961, 1994.

[19] W. Leland, M. Taqqu, M. Willinger, and D. Wilson. On the self-similar nature of Ethernet traffic (extended version). *IEEE/ACM Transactions on Networking*, 2(1):1–15, 1994.

[20] F. Masetti, P. Gavignet-Morin, D. Chiaroni, and G. Da Loura. Fiber delay lines optical buffer for ATM photonic switching applications. In *Proceedings of IEEE Infocom*, volume 3, pages 935–942, 1993.

[21] Y. Mei and C. Qiao. Effcient distributed control protocols for WDM optical networks. In *Proc. Int'l Conference on Computer Communication and Networks*, pages 150–153, September 1997.

[22] B. Mukherjee, S. Ramamurphy, D. Banerjee, and A. Mukherjee. Some principles for designing a wide-area optical networks. In *Proceedings of IEEE Infocom*, pages 110–118, 1994.

[23] P. Newman, G. Monshall, and T. Lyon. IP switching – ATM under IP. *IEEE/ACM Transactions on Networking*, 6:117–129, April 1998.

[24] V. Paxon and S. Floyd. Wide area traffic: the failure of Poisson modeling. *IEEE/ACM Transactions on Networking*, 3(3):226–244, 1995.

[25] C. Qiao and Y. Mei. On the multiplexing degree required to embed permutations in a class of interconnection networks. In *Proceedings of the IEEE Symp. High Performance Computer Architecture*, pages 118–129, February 1996. (A comprehensive version is to appear in IEEE/ACM Trans. on Networking).

[26] C. Qiao and Y. Mei. Wavelength reservation under distributed control. In *IEEE/LEOS Broadband Optical Networks*, August 1996. (a comprehensive version has been submitted to IEEE/ACM Transactions on Networking).

[27] C. Qiao, Y. Mei, M. Yoo, and X. Zhang. Polymorphic control for cost-effective design of optical networks. In *NSF DIMACS Workshop on Multichannel Optical Networks: Theory and Practice*, March 1998.

[28] C. Qiao and R. Melhem. Reducing communication latency with path multiplexing in optically interconnected multiprocessor systems. *IEEE Transactions on Parallel and Distributed Systems*, 8(2):97–108, 1997. (A preliminary version appeared in HPCA'96).

[29] C. Qiao, X. Zhang, and L. Zhou. Scheduling all-to-all connections in WDM rings. In *SPIE Proceedings, All Optical Communication Systems: Architecture, Control and Network Issues*,

volume 2919, pages 218–229, November 1996. (a modified version has been accepted for publication in IEEE/ACM Transaction on Networking).

[30] R. Ramaswami and A. Segall. Distributed network control for wavelength routed optical networks. In *Proceedings of IEEE Infocom*, pages 138–147, March 1996.

[31] R. Ramaswami and K.N. Sivarajan. Optimal routing and wavelength assignment in all-optical networks. In *Proceedings of IEEE Infocom*, pages 970–979, June 1994.

[32] R. Ramaswami and K.N. Sivarajan. Design of logical topologies for wavelnegth-routed optical networks. *IEEE JSAC/JLT Special Issue on Optical Networks*, 14(5):840–851, 1996.

[33] Y. Rekhter et al. Tag switching architecture overview. *IEEE Proceedings*, 82:1973–1983, December 1997.

[34] George N. Rouskas and Vijay Sivaraman. Packet scheduling in broadcast WDM networks with arbitrary transceivers tuning latencies. *IEEE/ACM Transactions on Networking*, 5(3):359–370, 1997.

[35] H. Shimonishi, T. Takine, M. Murata, and H. Miyahara. Performance analysis of fast reservation protocol with generalized bandwidth reservation method. In *Proceedings of IEEE Infocom*, volume 2, pages 758–767, 1996.

[36] Jonathan S. Turner. Managing bandwidth in ATM networks with bursty traffic. *IEEE Network*, pages 50–58, September 1992.

[37] E. Varvarigos and V. Sharma. The ready-to-go virtual circuit protocol : A loss-free protocol for multigigabit networks using FIFO buffers. *IEEE/ACM Transactions on Networking*, 5(5):705–718, October 1997.

[38] I. Widjaja. Performance analysis of burst admission-control protocols. *IEE proceedings-communications*, 142(1):7–14, February 1995.

[39] M. Yoo, M. Jeong, and C. Qiao. A high-speed protocol for bursty traffic in optical networks. In *SPIE Proceedings, All Optical Communication Systems: Architecture, Control and Network Issues*, volume 3230, pages 79–90, November 1997.

[40] M. Yoo and C. Qiao. Just-enough-time(JET): a high speed protocol for bursty traffic in optical networks. In *IEEE/LEOS Technologies for a Global Information Infrastructure*, August 1997.

[41] M. Yoo and C. Qiao. A new optical burst switching protocol for supporting quality of service. In *SPIE Proceedings, All Optical Networking: Architecture, Control and Management Issues*, volume 3531, pages 396–405, November 1998.

[42] X. Yuan, R. Melhem, and R. Gupta. Distributed path reservation algorithms for multiplexed all-optical interconnection networks. In *Proceedings of the IEEE Symp. High Performance Computer Architecture*, pages 38–47, February 1997.