Photonic Label Switch Supporting Asynchronous and Variable-Length Packets

Masayuki Murata, IEEE Member* and Ken-ichi Kitayama, IEEE Senior Member**

* Cybermedia Center, Osaka University
Toyonaka, Osaka 560-0043, Japan
murata@ics.es.osaka-u.ac.jp
** Graduate School of Engineering, Osaka University
Suita, Osaka 567-0871, Japan
kitayama@comm.eng.osaka-u.ac.jp

Abstract

In this paper, we propose new optical switching architectures supporting asynchronous and variable-length packets. Output line contention is resolved by optical delay line buffers. By introducing a WDM technology, parallel buffer can be equipped with multiple wavelengths on the optical delay line buffer. Differently from existing approaches, a main feature of our proposed architecture is that by employing a ultra-fast photonic label processing technique, an implementation is realistic for packet scheduling, which selects the appropriate output port, wavelength and delay line buffer. For evaluating the switch performance, an output port of our proposed switch is modeled as a multi-server and multi-queue system where each server corresponds to the wavelength. The arriving packet joins the shortest queue according to the scheduling policy of our switch. The switch performance is studied by utilizing an approximate analytic approach. Through the analysis and simulation experiments, we show that the introduction of the WDM technique can much improve the switch performance in terms of packet loss probabilities.

1. Introduction

The photonic network technology is expected to provide an infrastructure of the next-generation Internet against an explosive growth of traffic demands. In this section, a current status on the developments of the photonic technologies for carrying IP traffic is briefly reviewed. We also describe the background why we need an optical switch supporting asynchronous and variable-length packets.

The recent active development of the photonic technology increases the network bandwidth. Actually, we have already had a commercially available product utilizing the WDM technology to increase the link capacity between two adjacent routers. That is, each wavelength on the fiber is treated as a physical link between the conventional IP routers. In this way, the link capacity is certainly increased by the number of wavelengths multiplexed on the fiber. It is, however, insufficient to resolve a network bottleneck against an explosion of traffic demands since it only results in that the bottleneck is shifted to an electronic router.

There are two critical issues to be addressed in order to realize an ultra-fast optical packet switch that allows handling variable-length, asynchronous packets. These issues are originated from two major bottlenecks to the switch performance; time taken to look up the next-hop in the forwarding table, in which an incoming packet is matched with every entry in the table, and packet buffering for contention resolution, currently limited by DRAM or SRAM access times.

The first bottleneck in electronic routers is the longest prefix match for each incoming packet. The speed of a lookup algorithm of the routing table is determined by the number of memory accesses in order to find the matching entry, and the speed of the memory. The memory access time typically ranges from 10ns to 60ns. If an algorithm performs eight memory-lookups with a memory access time of 10ns, 12.5 million lookups/s can be performed [Kes98]. The bit rate of the link interface with the router is, for example, only 10Gbps for a 800-bit long packet. This processing capacity is much smaller than the aggregate capacity of WDM links of even one optical fiber ever achieved, e.g., 40Gbps x 160 wavelengths = 6.4Tbps/fiber

[Ito00]. It means that rough estimation that switching speeds are 20 times greater than forwarding speeds for comparably priced hardware [Lin97] still holds even for photonic technologies.

One promising way to alleviate the capacity limit of the routers is to introduce an MPLS (Multi-Protocol Label Switching) technology [Dav98]. By MPLS, switching and forwarding capabilities are separated to fully utilize a high speed switching capability of the underlying network such as ATM. Packet forwarding to determine the destination is only performed at the edge of the MPLS domain. While MPLS needs to establish a closed domain for utilizing a new lower-layer technology, it is useful to incorporate the photonic technology for building the very high-speed Internet. However, there are still several problems in order to deploy λ–MPLS. The most difficult problem in λ–MPLS is a capacity granularity; the unit of the bandwidth between edge node pairs of the MPLS domain is a wavelength capacity. It may be sometimes too large to accommodate the traffic between node pairs. One approach to resolving the capacity granularity problem is addressed in [Ban00], where the authors introduce wavelength merging, but the related technology is still immature.

Another promising technique is to utilize a photonic label switching technology recently developed in [Kit99] where the above granularity problem can be resolved in an optical domain, thus improving bandwidth efficiency. As shown in Figure 1, a photonic label is attached to the head of the payload data. A family of optical code sequences is utilized as the photonic labels, which has been originally used as signature codes in the optical code division multiplexing (OCDM) [Pru86], [Sal89]. The recognition of the optical code among the codes is accomplished based upon optical correlation in the optical domain. As the optical correlation can be performed simply by using passive optical waveguide device, the recognition time is governed by the propagation delay of the device. This is a key to the ultra-fast photonic label processing [Kit99]. Hereafter, we will call the MPLS technology utilizing the photonic label switching method as *OC-MPLS*. Its ultra-fast photonic label processing capability is expected to be suitable for ultra-high bit rate MPLS applications.

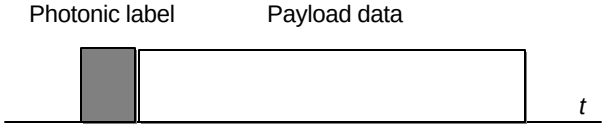

Figure 1: A Structure of Optical Packet Attached with a Photonic Label

The second bottleneck is caused by queuing and buffering to resolve the contention at the output ports of the switch. We need to develop an optical packet switch resolving the contention at the output ports, which is our main subject of this paper. For the packets failing to acquire the output line, we add the fiber delay line for buffering packets. By sharing the optical delay line buffer by WDM, a performance improvement can be expected. For that purpose, however, we need to introduce the scheduling algorithm for putting the packet to the appropriate queue. Fortunately, the photonic label switching has an capability to process the packet at very high speed within the packet switch. Supporting IP packets directly can also be effectively implemented by the photonic label switching. That is, our design allows asynchronously arriving packets with variable length.

From an architectural point of view, our proposed concept is similar to ATM switches and ATM-based MPLS. However, ATM only allows packets with fixed size (called *cells*). This is a very different point since our architecture does not introduce an excessive overhead of packet assembly/disassembly at the edge router of MPLS (Ingress LSR).

As related works, active research efforts have been recently made on optical packet switches. However, most of those researches have been devoted to the switches handling fixed size packets. See, e.g., [Hun98]. Only a few exceptions include the approach described in [Tan00], where the authors propose the optical packet switch supporting asynchronous and variable-length packets. The scheduling algorithm based on a concept of a void filling is also proposed in order to resolve output contention at the delay line buffer. The authors in [Ge00] consider the scheduling policy for storing simultaneously arriving packets into the optical buffer with different wavelengths. They compare four scheduling policies in terms of packet loss probability. A main problem of the existing approaches is that the packet header processing is assumed to be performed in an electronic domain, and henceforth complicated scheduling policies described in the above literature is likely to become a bottleneck. On the contrary, we carefully consider the implementation issues for realizing a very high-speed packet switching in our proposed switch.

This paper is organized as follows. In Section 2, we will describe our switch architecture to handle asynchronous and variable-length packets in an optical domain. Implementation issues for the switch are discussed in detail in Section 3. In Section 4, the performance of our proposed switch is evaluated through both of approximate analysis and simulation experiments. Section 5 concludes our paper

with future research topics.

2. Switch Architecture

We will propose two switch architectures without and with wavelength conversion. If wavelength conversions are not allowed, the hardware becomes simple. On the other hand, the operation of the switch with wavelength conversions slightly becomes complicated, but we can attain high performance as will be demonstrated in Section 4. In this section, we will introduce our optical switch architectures, and the implementation issues are described in the next section.

2.1. Optical Switch without Wavelength Conversion

We start describing the optical switch without wavelength conversion. It consists of three sections: optical switching unit, optical scheduling unit, and optical buffering unit. See Figure 2, in which a 2x2 optical switch is illustrated. The number of wavelengths is *W*. As shown in the figure, each component is dedicated to a single wavelength channel.

The optical switching unit switches packets according to the photonic label information. If the packet is destined for the output port O_1, then the switch is set to the bar-state, directing the packet to the upper part of the optical switching unit. It can be performed by the photonic label processing, and the label recognition time is only governed the propagation delay of the optical decoder device, leading to the ultra-fast photonic label processing [Kit99]. It has been predicted based upon the experimental results [Wad00] that the processing speed of the photonic label more than 10^9 packet/s can be attained.

The optical buffering unit provides an optical buffer by using fiber delay lines. Let *D* be a delay line unit. Then, to delay the packet during *iD*, it is put on the *i*-th delay line (shown by τ_i in the figure). The counter b_{ij} keeps the buffer status information for wavelength λ_j going to the output port O_i. To handle the variable-length packets, it is incremented when the packet arrives at the optical buffer as

$$b_{ij} \leftarrow b_{ij} + \lceil x / D \rceil \qquad (1)$$

where *x* denotes the length of the arriving packet. It is decremented by one for every *D* time unit. Then, the next arriving packet is put on the b_{ij}-th delay line.

The heart of our optical switch is the optical scheduling unit. Without wavelength conversion in Figure 2, each op-

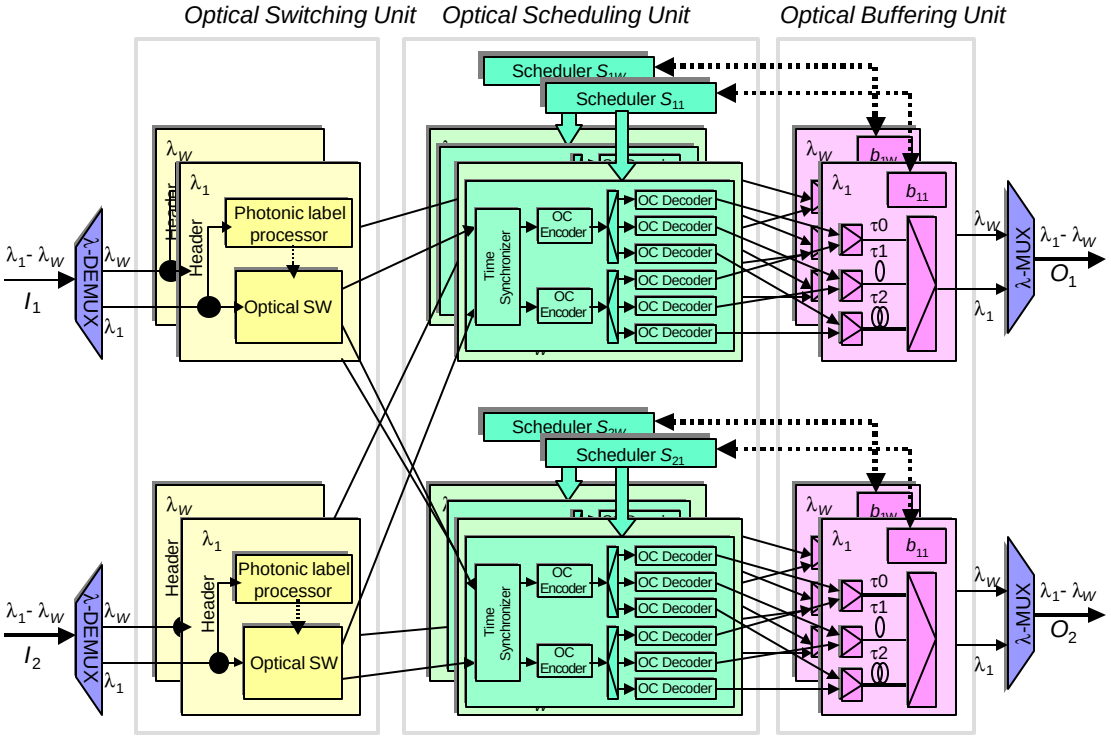

Figure 2: Optical Switch without Wavelength Changes

tical scheduler S_{ij} is dedicated to schedule packets on wavelength λ_j destined for the output port O_i. It reads the packet length in the header of the arriving packet, and updates the buffer status b_{ij} according to Eq. (1). Then, after the time synchronization each bit of the payload in the optical packet is encoded with an optical code, which is of the same structure as the photonic label in Figure 1, used outside the photonic label switch. The optical code indicates a specific delay line in which the packet is to be fed. The encoded packet is split and delivered to the optical decoders. Only one of the decoders recovers the data bits of the packet, , and then the packet is transferred to an appropriate delay line. See Subsection 3.1 for the optical buffer assignment based on photonic label processing.

One problem is that we need to handle simultaneously arriving packets from different input ports. Each scheduler performs the three-step operation for each packet exclusively for maintaining the valid counter: (1) to read the buffer status information, (2) to update it according to the packet length, and (3) to write it back onto the memory. It must not receive another packet during the three-step operation. A time synchronizer is introduced for this purpose. A role of the time synchronizer is to delay another packet if the scheduler processes the packet. Since only a single packet arrives from an input port at a time, it is sufficient that a time synchronizer is prepared for each input port. The hardware complexity increases by introducing the time synchronization. All incoming packets destined for the output port O_i should be processed in sequence. See Subsection 3.3 for more detail.

2.2. Optical Switch with Wavelength Conversions

We next consider the switch with wavelength conversion. See Figure 3. The wavelength converters, indicated by supercontinuum light source with gate switches (SC + Gate), are added in the optical scheduling unit. The wavelength conversion allows incorporating the WDM buffer into the switch fabric. In this case, the packet in contention will be put on the alternate delay line buffer by changing the wavelength in front of the buffer. See Subsection 3.2 for a novel wavelength conversion method. By introducing the wavelength conversion, a significant reduction of the packet loss probability is expected, which will be discussed in more detail in Section 4.

The optical scheduling unit of the switch allowing wavelength conversion slightly becomes complicated. To schedule the packets destined for the output port O_i, the scheduler S_i of the output port O_i has to know the status of

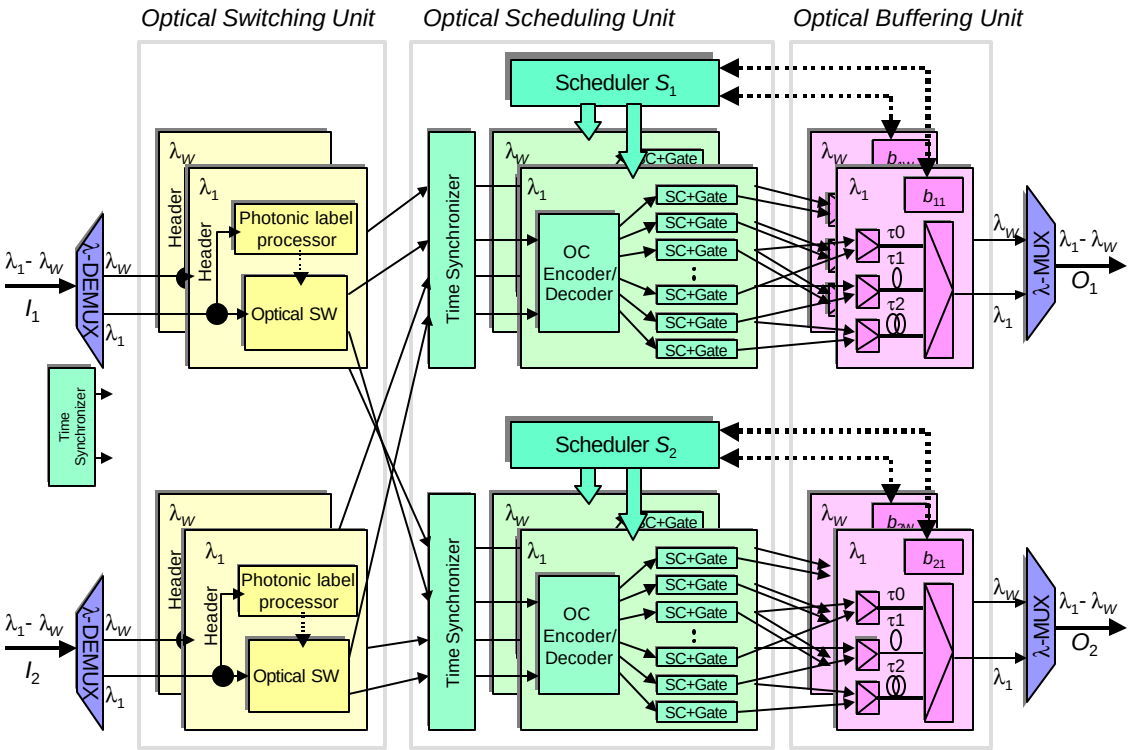

Figure 3: Optical Switch with Wavelength Conversion

three delay lines τ_0, τ_1, and τ_2 with W different wavelengths. When the packet arrives, the scheduler searches the shortest queue. It can be easily determined by the counters b_{i1} through b_{iW}. A simple comparator in hardware may implement it. Once the scheduler determines the wavelength to be tuned (say, λ_j), it updates the counter b_{ij} and sets the gate for selecting the wavelength.

3. Optical Implementation Issues

3.1 Optical Buffer Assignment Based upon Photonic Label Processing

Photonic label processing is also exploited within the switch for buffer assignment. Optical codes used in the optical buffer assignment are the same as those used as the photonic labels for network-wide OC-MPLS outside the photonic switch. Distinct from the photonic labels used in OC-MPLS is that it does not require global significance in the network but local significance within the switch. Therefore, the scheduler has to be provided only with the information of the buffer status within the switch. Figure 4 illustrates the schematic of proposed buffer assignment in the optical domain. For simplicity, suppose that the buffer has three delay lines, τ_0, τ_1 and τ_2, with different lengths as shown in Figure 4. The packet from the output port of the switch is encoded with a photonic label which designates an appropriate delay line to be fed and is delivered to all the optical decoders 1 through 3. The scheduler determines the appropriate delay line for the packet to be fed, and assigns a photonic label. By encoding the packet with the photonic label, the output emerges only from the decoder, which is assigned with the same label as that of the incoming packet. Note that each bit of the payload is encoded with the photonic label, while in OC-MPLS, the photonic label is used only in the header as shown in Figure 1, and the payload data bit is unencoded. The output from the decoder recovers an original bit sequence of the payload, and the recovered packet is fed into a desired delay line, resulting in the contention resolution in the optical domain.

Figure 5 shows a special class of optical encoder/decoders [Wad99]. The incoming pulse stream (from the *l.h.s* on the top) is encoded into 8-chip bipolar phase-shift keying (BPSK) optical code through the tapped delay line, followed by the optical phase shifter (from the *r.h.s* on the bottom). The optical carrier of the split pulse is phase-shifted by 0 or π, resulting in the bipolar phase-shift

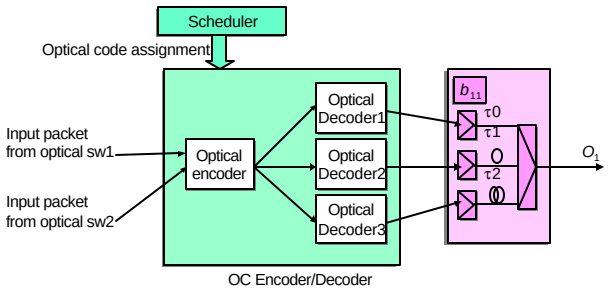

Figure 4: Buffer Assignment in Optical Domain

keying. Both the tapped delay line and the phase shifter are tunable, retaining the programmability of optical codes. The decoding is carried out with the same device. The optical code launched from the *r.h.s.* on the top is correlated, and the auto-correlation peak emerges if the assigned code matches that of the incoming optical code (from the *l.h.s* on the bottom). Obviously, the time optical correlation takes is equal to the propagation time of the code in the decoder. In the decoder of Figure 5, it takes only 70ps for the chip pulse at the tail to pass through after the leading chip pulse enters the device. This is the key to the ultrafast photonic label processing capability. It is noteworthy that all the process is carried out in the optical domain, and the photonic label recognition is performed without any logic operation. If we use the number of delay lines (i.e., the buffer size) to be 10, the number of chips of 5 is large enough for the optical codes, by which 2^4 codes are available. Actually, the buffer size of 10 is a reasonable design choice, as we will discuss in Subsection 4.4.

Based upon the experimental results of the photonic label recognition using 8-chip BPSK optical codes has been successfully demonstrated at the bit rate of 10Gbps [Wad00], it is predicted that the bit rate can go up to 100Gb/s for 5-chip BPSK optical code using readily available optical pulse with 1ps pulsewidth. Compared with the optical code in Figure 5, the chip pulse interval is shrinked from 5ps to 1ps.

3.2 Super-continuum Wavelength Converter

In this subsection, we will focus on the switch architecture in Figure 3 and describe devices for the wavelength conversion between the optical decoder and the optical delay line buffer. To share the delay line with packets of different wavelengths, WDM buffering can be adopted. For example, when the packet from the decoder on λ_1 plane fails to find an empty delay line in b_{11} but it finds a non-

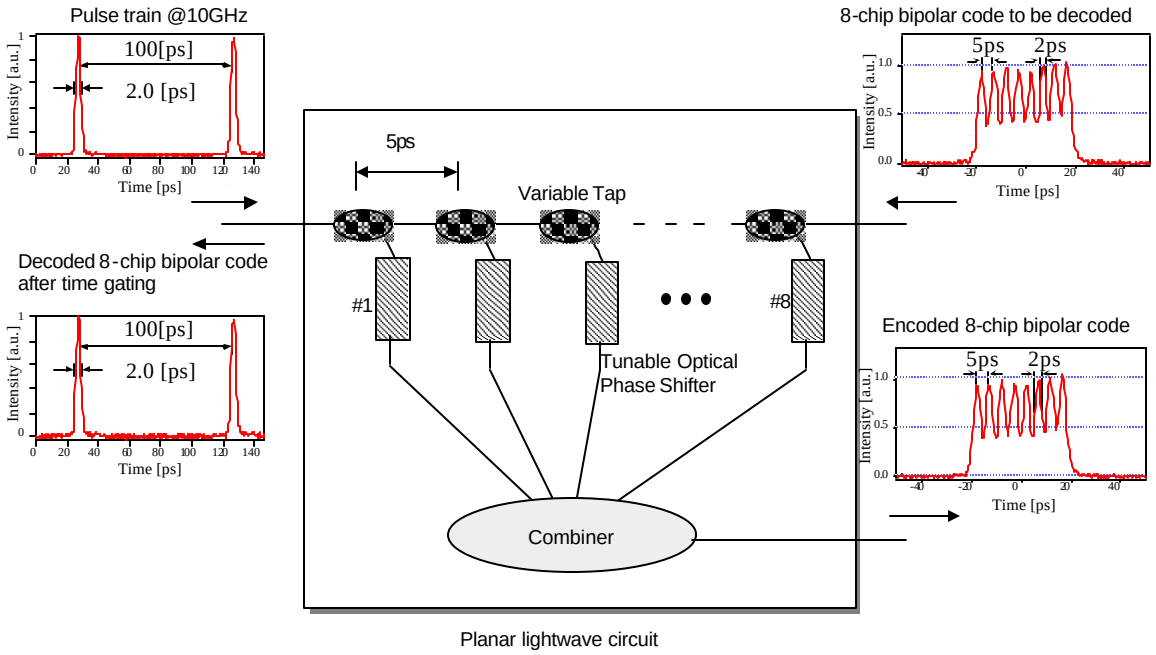

Figure 5: Optical Encoder/Decoder

empty delay line τ_0 on b_{1k}, it can be fed to the nonempty delay line after converting the wavelength of the packet to λ_k. Note that the group delay time difference caused by the difference of the wavelength is negligible.

We propose a future scheme of wavelength conversion, rather than readily available semiconductor optical amplifier (SOA) tunable wavelength converter [Dan98]. Although, we must not neglect the practical availability of SOA wavelength converters, we will rather focus on a novel device that is preferably applied to ultra-high bit rate dense WDM (DWDM) systems in the future. As shown in Figure 6, it consists of a supercontinuum (SC) fiber, combined with a wavelength demultiplexer (AWG) and an SOA gate switch array (Gate switch). Compared with conventional SOA wavelength converter, advantages include the ultra-fast response of SC due to the nature of the fiber nonlinearity and the ultra-wide spectral region of the emission, high-speed gate switching and a large count of the output ports of AWG. However, its larger number of components by a factor of $3W$ where W is the number of WDM channels, might be costly, and additional optical amplifiers are required to have the input signal serve as the pump.

A supercontinuum (SC) pulse source generates picosecond pulses at several tens of Gbps over an extremely broad spectral range. The seeding short pulse at a specific repetition rate is broadened continuously in its spectrum due to

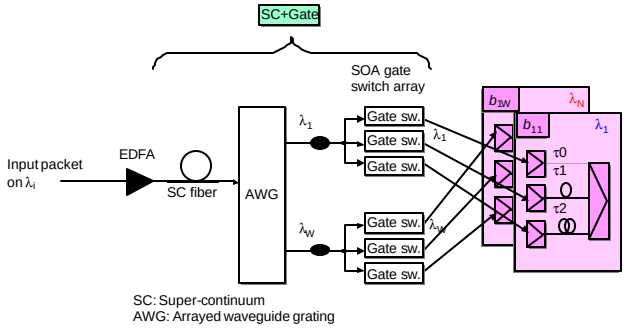

Figure 6: Structure of Wavelength Conversion Unit

the fiber nonlinearlities, and thus by filtering desired wavelength components out of the SC spectrum by AWG, the single light source can serve a multi-wavelength pulse source. It is another advantage that it uses only one pump laser, and its fixed channel spacing with accuracy equivalent to that of a microwave oscillator (Hz), enabling to lock the entire chain to absolute standard by locking just one mode of the chain. In Figure 7, a typical SC spectrum, the temporal waveform and the spectra of the spectrum-sliced output are shown [Sob98]. Recently, 1,010-channel with the frequency interval of 12.5GHz at 2.5Gb it/s over 100nm in 1550nm spectral region has been achieved [Tak00].

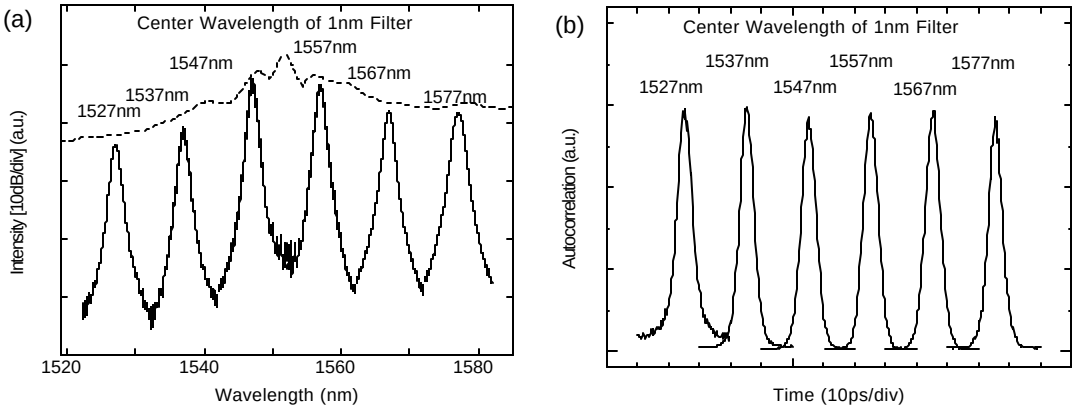

Figure 7: Typical Supercontinuum (SC) Spectrum at the Repetition Rate of 10Ghz; (a) Optical Spectra of Spectrum-Sliced WDM Channels and (b) Their Temporal Waveforms. (Reproduction from [Sob98])

3.3 Time Synchronization

It should be guaranteed that a packet does not enter the optical switching unit before the scheduler completes its task to direct the preceding packet to the buffer. A time synchronizer is introduced for this purpose. We will impose a constraint on the incoming packets that the incoming packets have to be waited at the input port until the scheduler of the input port completes the task for the preceding packet. To materialize this constraint, the time-synchronizer, placed in the front of the optical OC encoder in the optical scheduling unit, aligns the incoming packet to the time-slot as shown in Figure 8. Note that the synchronization is local within the switch and is independent of the global clock that the electrical interface of the switch has, thus maintaining asynchronous operation at the bit level. It is also noted that the slot duration is set equal to the processing time of the scheduler for a packet, and the time duration is much longer than one-bit duration of the clock. For example, assume that the processing time of the electronic scheduler is roughly 2ns, and the time resolution of the synchronization has to be less than 10% of the time slot, then the maximum delay time required for the time synchronizer becomes 2ns with the time resolution of less than 0.2ns.

The block diagram of the time synchronization is depicted in Figure 9. The operation of time synchronization includes time alignment along with start recognition and time evaluation. The start time of the packet is identified from the packet header, and the necessary delay time is calculated. Only the time alignment has to be carried out in the optical domain while the start recognition and time evaluation can rely on the well-established electronics to

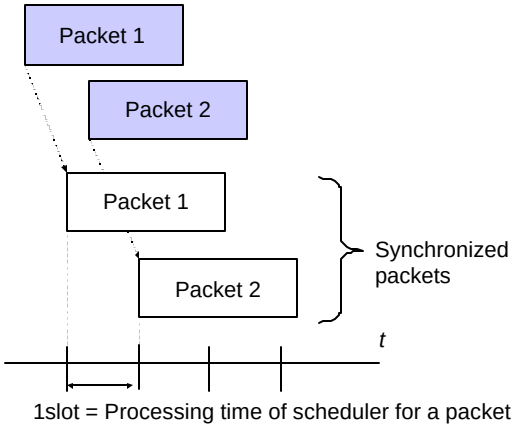

Figure 8: Time Synchronization among Packets

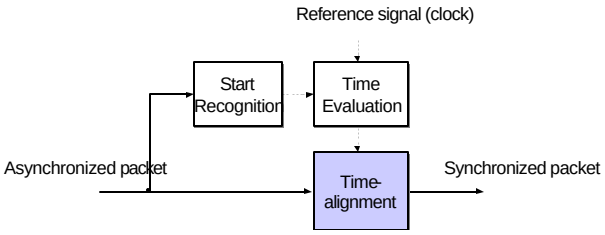

Figure 9: Block Diagram of Time Synchronization

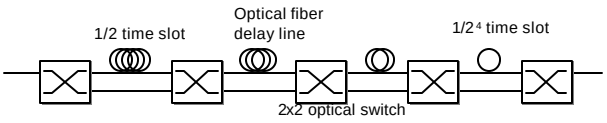

Figure 10: Cascaded Delay Line Buffer

meet the operation speed. The time alignment can be realized by using a variable delay line. There have been several synchronization schemes for "coarse" and "fine" time alignments, respectively, using a switched delay line [Chl96] and group delay dispersion combined with the wavelength conversion [Fra99]. A suitable device would be the cascaded switched delay line shown in Figure 10. The lengths of the delay lines are arranged so that the first delay line is equal to 1/2 time slot duration, the second one is equal to $1/2^2$ time slot duration, and so on. Each 2x2 optical switch between the delay lines is configured with the correct path based upon the time evaluation. Four-cascaded delay line with the first delay line of 200mm-long corresponding of 1ns can guarantee the time resolution of less 0.2ns.

4. Performance of Proposed Optical Switch

In this section, we evaluate our optical switch proposed in the previous sections. Our analysis is rather simple and, for example, we do not consider the time synchronizer described in the previous section. However, it also implies that our analysis approach and results obtained in this section are generic, and our discussion here can be applied to the other optical switches utilizing the fiber delay lines as a packet buffer.

4.1. Analysis Approach

To evaluate our optical switch, we focus on one output port since the switch is non-blocking. In the switch without wavelength conversion, packet switching is performed on each wavelength independently. The packets arriving at the input ports are switched to the designated output port in the optical switching section. Thus, it can be modeled as a single queue, where the server corresponds to one wavelength. In this case, our concern is the influence of introducing the fiber delay line as the packet buffer, which is studied in [Cal00]. When the packet arrives and the server is idle, the packet is transmitted immediately. If the server is busy, on the other hand, the packet is queued. Let t be the arrival time and t_f be the time at which the server will be free to serve the new packet. The new packet can be served after $t_f - t$ in the case of electronic buffers. However, in the current case, we need to consider the "granularity" of the fiber delay lines. That is, when the fiber delay line buffers the packet, the buffering time is measured by the delay unit of the fiber delay line, D, and therefore, only a finite set of delays can be achieved. The new packet is delayed by an amount of

$$\Delta = \left\lceil \frac{t_f - t}{D} \right\rceil D \qquad (2)$$

In other words, an excess length of $\Delta - t_f + t$ is additionally brought to the server in the case of fiber delay line buffers.

Based on the above observation, the author in [Cal00] introduces an additional service time for each packet if the packet finds the server to be busy. In the queueing system, which can be described by a birth-and-death process, the author presents an iteration algorithm to find the packet loss probability. The granularity D affects the performance as follows; if D is small, the time resolution of the fiber delay line is small and therefore the performance must improve with decreasing D, but the buffer capacity in bytes becomes small. If D becomes large, the buffering capacity gets large since the long delay is introduced. However, at the same time, the time resolution of the buffer becomes small and the larger excess load is introduced. Through numerical examples, the author finds that there exists an optimal value of D around 0.3 irrespective of the number of the buffer size when the average packet size is one. We note here that the number of delay lines corresponds to the buffer size, which is represented by B. The buffer capacity in bits is then determined as $B \times D$.

For the switch allowing wavelength conversion, we have multiple servers, each of which has a dedicated queue. See Figure 11. Our scheduling policy is to place the packet at the shortest queue if all servers are busy. In our switch, it is achieved by choosing the smallest counter. This sort of the

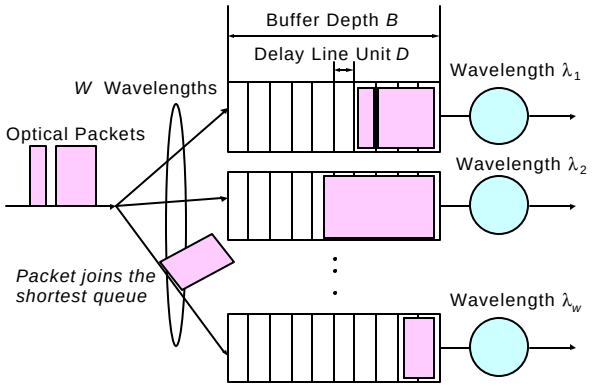

Figure 11: Model for Optical Delay Line Buffer utilizing WDM

system is studied in [Lin96], in which the authors studied a multi-server and multi-queue system with a "Join the Shortest Queue" (JSQ) policy. In our case, the server corresponds to the wavelength, and each has a buffer by optical delay lines.

An approximate analysis of our system consists of two parts. The first one is to extend the approximate analysis developed in [Lin96]. Since in the original approach, the buffer capacity is assumed to be infinite, we introduce another approximation to treat the finite buffer capacity. The second part is to apply the approach in [Cal00] to take account of the granularity of the fiber delay lines. Since our approximate analysis is rather straightforward, we summarize our approach in Appendices A and B.

4.2 Effects of Multiple Wavelengths

In this subsection, we provide numerical results by applying our analysis. The traffic load ρ per wavelength is given by

$$r = l/(mW) \qquad (3)$$

where λ is a total packet arrival rate at the output port and μ is an inverse of the average packet length. In numerical examples, the packet arrival rate λ is set to be in proportion to the number of wavelengths *W*, and the traffic load per wavelength is fixed. For simple presentation, we set the average packet length to be unity, and set *D* to be relative to the average packet length.

The first result shows an effect of the delay line unit *D* in Figure 12. Six values of the number of wavelengths are considered: *W* = 1, 2, 3, 4, 6 and 8. The analytical results are plotted with solid lines. For *W* = 1, 2, 3 and 4, the simulation results are also presented to assess the accuracy of our approximate analysis. In simulation, we generated a billion packets. The traffic load ρ is fixed at 0.8 and the buffer depth *B* is set to be 64. Note that the case of *W* = 1 corresponds to the result provided in [Cal00]. From the figure, it is clear that the optimal value of *D* is around 0.3 irrespective of the number of wavelengths. Another observation is that the increasing number of wavelengths can dramatically improve the switch performance if *D* is appropriately selected. In the current example setting, the traffic load per wavelength is identically set. It implies that the case of *W* = 1 corresponds to the switch without wavelength conversion, in which each of multiple wavelengths forms an independent queue. In the case of *W* = 1, the packet loss probabilities are very high, and still more the result is rather optimistic because it implicitly assumes that

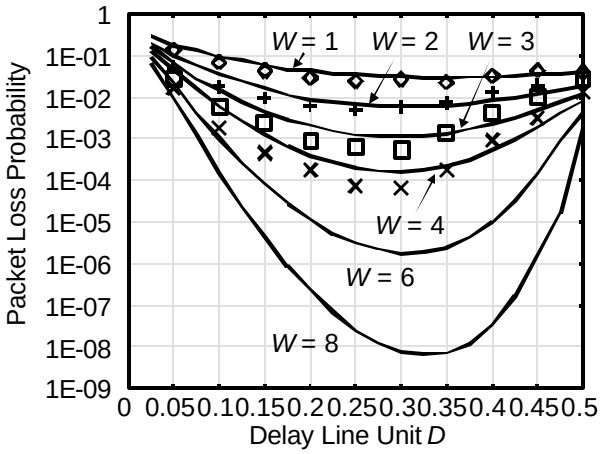

Figure 12: Packet Loss Probabilities dependent on *D*

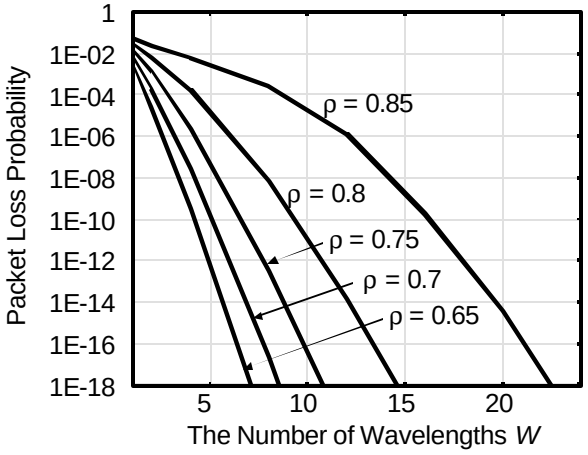

Figure 13: Packet Loss Probabilities dependent on the Number of Wavelengths *W*

the traffic load is well balanced among independent wavelength channels of *W*.

To clearly see the effect of introducing the wavelength conversion in the switch, we next plot the packet loss probabilities dependent on the number of wavelengths in Figure 13. Five values of the traffic load are used: ρ = 0.65, 0.7, 0.75, 0.8 and 0.85. The buffer size is set to be 64. The effect is apparent.

Another view of the effect by the number of the wavelengths is the next presented for buffer dimensioning. Preparing the number *B* of delay lines for each output port directly affects the switch cost. Thus, the buffer size is an important design parameter. Figures 14 and 15 show the packet loss probabilities dependent on *B* for *W* = 1 and *W* = 8, respectively. As can be observed in Figure 14, a quite large amount of the buffer is necessary to decrease the packet loss probabilities in the switch without wavelength conversion. By increasing the number of wavelengths *W*

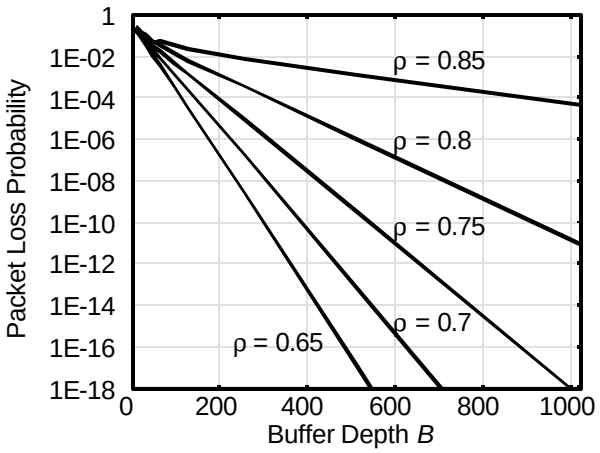

Figure 14: Packet Loss Probabilities dependent on Buffer Depth B (W = 1)

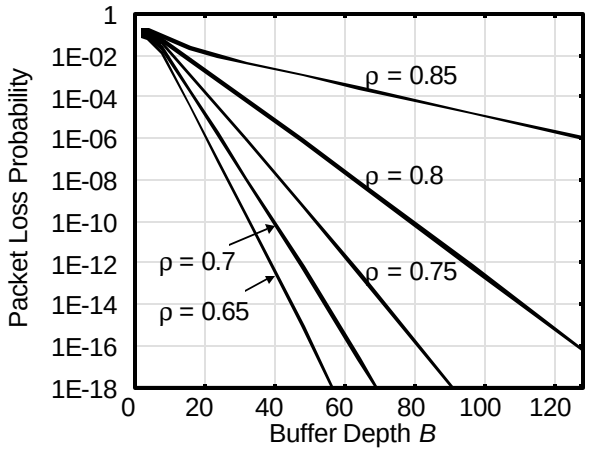

Figure 15: Packet Loss Probabilities dependent on Buffer Depth B (W = 8)

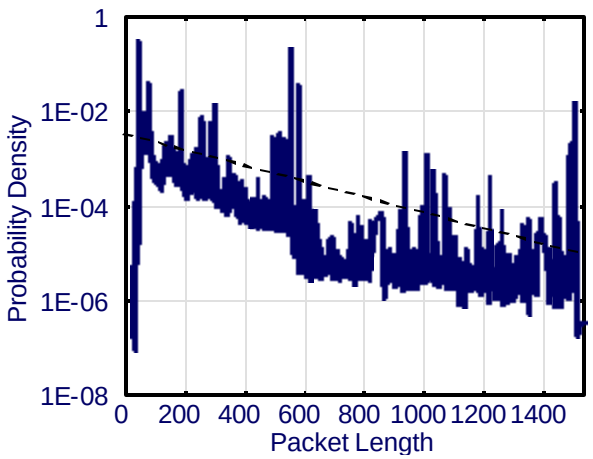

Figure 16: Packet Size Distribution from [WAN97]

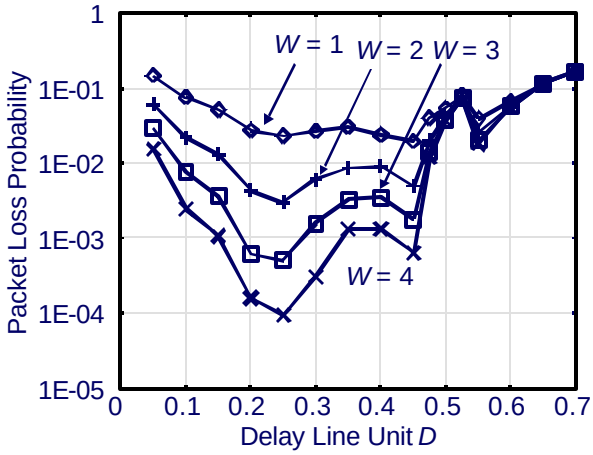

Figure 17: Packet Loss Probabilities dependent on D (Packet size distribution of Figure 16 is used.)

from 1 to 8, almost one-tenth of buffer capacity is sufficient to attain the same packet loss probabilities for given traffic load per wavelength.

4.3 Influences of Packet Size Distribution

Since the Internet packet size does not follow an exponential distribution, we next investigate its influence on the switch performance. For this purpose, we used the actual traced data found at [WAN97]. See Figure 16, in which the exponential distribution with same mean is also plotted for reference purpose. Note that in the figure, we omit very small probabilities for the packet with about 4,000 bytes. Since it is difficult to evaluate the switch performance though an analytic approach, we conducted simulation experiments. For consistency, we set the average packet size to be unity (corresponding to 257.1 bytes in the current traffic data). Since we used the simulation, the parameter region for the small packet loss probabilities could not be examined. The results are shown in Figure 17. A difference from Figure 12 (for the case of exponentially distributed packets) seems to be caused by the packet size distribution. In the current case, the probabilities around 500-byte packet are large as shown in Figure 16, leading to the large packet loss probabilities around D = 0.5. Except for the difference, it can be observed that the effect of the packet size distribution is not large, and we can expect that the analysis results presented in the previous subsection hold even for the actual packet size distribution with very small packet loss probabilities. It is especially true that the delay line unit of D is set to be around 0.25 gives best result.

4.4 Optimizing the Number of Wavelength and the Buffer Size

We last discuss the optimization of the WDM buffer. There will be optimum values of the number of the wavelengths W and the number of delay lines, that is, the buffer size B, to satisfy a requirement for the packet loss probability. One problem with the WDM buffer of a large size would be a skew, which is the group delay difference between the optical fiber delay lines caused by the group delay dispersion of the fibers. When the packets of different wavelengths are fed into a delay line, two or more packets emerge at different timing, making the administration by the schedulers complicated. In Table 1, the group delay differences for various buffer sizes against the number of WDM channels are summarized. Here, we assume that the WDM channel spacing is 1nm, and a standard single-mode fiber, having the dispersion value of 17ps/nm/km at the center wavelength of λ = 1550nm, is used for the delay lines. Note that a unit length of the delay line is set to be equal to the average packet size of 2,000-bit at the bit rate of 100Gb/s and D=1.0 in calculation.

Table 1: Skew [ns] vs. The # of Wavelengths and Buffer Sizes

	Buffer Size B [in Packets]		
W	10	100	1,000
10-channel	$6.8\text{x}10^{-3}$	$6.8\text{x}10^{-2}$	$6.8\text{x}10^{-1}$
100-channel	$6.8\text{x}10^{-2}$	$6.8\text{x}10^{-1}$	6.8

As shown in Table 1 for W > 100 and B > 1,000, the skew becomes prohibitedly large. Considering the tolerable range of the skew will be within the 10% of the packet length, it becomes 0.4m. Note that 2,000-bit long packet at the bit rate of 100Gb/s is 4m. Therefore, 100 WDM channels with the buffer size of 1,000-packet may not be allowed. A long optical fiber delay line has also another problem that the each optical pulse of the packet data bit suffers its waveform distortion due to the dispersion effect and the optical loss. To compensate the waveform dispersion and the optical loss, dispersion compensation fibers and optical amplifiers [Hal99], respectively, have to be introduced. These make the buffer rather complicated.

Let us set a criterion that the allowable packet loss probability is 10^{-9} for the traffic load ρ = 0.8. From the numerical simulations in Figures 14 and 15, for W = 1 and B > 800, and B decreases to 70 for W = 8. From another viewpoint in Fig.13, W > 16 for B = 64. By taking these observations into consideration, it is concluded that the optimum range exists around W = 10 and B = 100. The buffer size of B = 100 corresponds to the fiber delay line length of 120m for the average packet size of 2,000 bits at the bit rate of 100Gb/s with D = 0.3. These numbers are practically feasible because commercial products of WDM systems with more than more than a hundred channels are around the corner , and the 120m-long optical fiber delay line could be constructed without using an optical fiber amplifier. It may be predicted that if the WDM buffer has more than a hundred of wavelengths, the packet loss probability becomes almost negligible.

5. Concluding Remarks

In this paper, we have proposed a new optical switch based on the photonic label switching techniques. Our switch can support asynchronous and variable-length packets. Contention resolution at the output buffer is resolved by introducing fiber delay lines as packet buffers. By incorporating the WDM technology into optical delay line buffer, the switch performance is dramatically improved. It has been shown that the WDM buffer assignment can be implemented using ultra-fast photonic label processing. This has been shown by a newly developed approximate analysis method. We have also tested the case of generally distributed packets. Another important case is related to the packet arrivals; i.e., a heavy-tailed distribution might be necessary for actual buffer dimensioning [Tan00]. However, it is still questionable whether such arrivals occur at the backbone switches that we are considering, and more researches are necessary in this field.

We have intended that our proposed switching architecture is applied to MPLS-based networks. Its advantage is that the packet loss can be well dimensioned by the traffic engineering approach [Dan00]. However, it requires building a closed cloud of MPLS networks. Another possibility is that our proposed switch works just as IP routers. Since the photonic label does not restrict on the label contents, the destination IP address can be used as a photonic label. A longest prefix matching is also allowed by the current photonic label processing technology [Kit99]. Furthermore, more flexible addressing can be utilized. For example, we can embed the source IP address and port numbers within the photonic label, by which flow discrimination and service quality for each flow can be controlled at the switch. For our switch acting as the IP routers, however, a table

update should be implemented because in the current technology, it requires a large time.

Appendix A: Analysis of Multi-Server Multi-Queue with "Join-the-Shortest-Queue" Policy

In this appendix, we develop an approximate analysis for our optical switch with wavelength conversion. The operation of our switch can be modeled by a multi-server and multi-queue system where each server is equipped with finite buffer, and a newly arriving job, which finds no idle servers, is buffered at the shortest queue. Here, the server corresponds to the wavelength, and the job does the variable-length packet.

Since the buffer is implemented by the fiber delay line, we need to consider the granularity of the delay unit, which affects the switch performance. Such a study can be found in [Cal00]. In [Cal00], however, the author considered the single queue governed by the birth and death process. That is, the single wavelength is treated in [Cal00].

On the contrary, our system has W wavelengths, and the delay line can be shared by those wavelengths. It leads to the above-mentioned multi-server and multi-queue system. In [Lin96], the authors treat such a system with infinite buffer capacity. We extend the analysis developed in [Lin96] to treat the finite buffer case. A key of the analysis in [Lin96] is that an evolution of the system behavior is represented by the simple birth-and-death process as shown in Figure A.1, where the total number of jobs in multiple queues is considered as a Markov state. In order to take account of the scheduling policy, the state-dependent service rates are considered. As shown in the below, the algorithm requires an iteration, and state-dependent service rates at i-th iteration is represented by $m_k^{(i)}$ $(k = 1,2,3...)$. Our modification is rather straightforward for treating the finite buffer, and therefore, we only show the results without explanation.

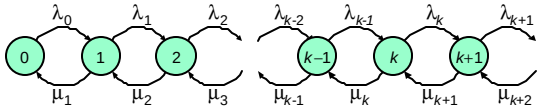

Figure A.1: State Transition Diagram

By letting μ be the packet transmission rate on each wavelength, the state-dependent service rates are determined by the following equations:

$$m_1^{(i+1)} = m \tag{4}$$

$$m_k^{(i+1)} = a_k(m_{k-1}^{(i)} + m) + b_k B_k, \ \ k = 1,\ldots,W-1 \tag{5}$$

$$m_k^{(i+1)} = a_k A_k + b_k B_k, \ \ W \le k \tag{6}$$

where

$$A_k = \begin{cases} \sum_{j=0}^{a_{k-1}} \binom{a_{k-1}}{j} \left(q_{k-1}\right)^j \left(1-q_{k-1}\right)^{a_{k-1}-j} \\ \qquad \times g_k(j+1), & \text{if } q_{k-1} < 1 \\ g_k(a_{k-1}+1), & \text{if } q_{k-1} = 1 \end{cases} \tag{7}$$

$$B_k = \begin{cases} \sum_{j=0}^{a_{k+1}} \binom{a_{k+1}}{j} \left(q_{k+1}\right)^j \left(1-q_{k+1}\right)^{a_{k+1}-j} \\ \qquad \times g_k(j+1), & \text{if } q_{k+1} < 1 \\ g_k(a_{k+1}+1), & \text{if } q_{k+1} = 1 \end{cases} \tag{8}$$

$$g_k(j) = \begin{cases} (j+1)m, & \text{if } j < W \\ W m, & \text{if } j = W \end{cases} \tag{9}$$

$$h_k(j) = (j-1)m(1-w_k(j)) + j m w_k(j) \tag{10}$$

$$w_k(j) = \begin{cases} (k+1-j)/j, & \text{if } k+1-j < j \\ 1, & \text{if } k+1-j \ge j \end{cases} \tag{11}$$

$$a_k = \min(k, W) - 1 \tag{12}$$

$$q_k = \frac{1}{a_k}\left(\frac{m_k^{(i)}}{m} - 1\right) \tag{13}$$

$$a_k = \frac{p_{k-1}^{(i)} l_{k-1}^{(i)}}{p_{k-1}^{(i)} l_{k-1}^{(i)} + p_{k+1}^{(i)} m_{k+1}^{(i)}} \tag{14}$$

$$b_k = \frac{p_{k+1}^{(i)} m_{k+1}^{(i)}}{p_{k-1}^{(i)} l_{k-1}^{(i)} + p_{k+1}^{(i)} m_{k+1}^{(i)}} \tag{15}$$

The steady-state probabilities for the birth-and-death process can be obtained as in a usual way:

$$p_k^{(i)} = p_0^{(i)} \prod_{j=0}^{k-1} \frac{l_j^{(i)}}{m_{j+1}^{(i)}}, \ \ k = 1,2,\ldots \tag{16}$$

$$p_0^{(i)} = \left[1 + \prod_{j=0}^{k-1} \frac{l_j^{(i)}}{m_{j+1}^{(i)}}\right]^{-1} \tag{17}$$

A main difference from the original approach can be found in Eqs. (14) through (17) where packet arrival rates are also state-dependent. In state k, the arrival rate depends on the buffer status because we should exclude the lost packets. Since an essence of the JSQ policy is to balance the buffer occupancy among multiple queues, we assume that jobs are evenly distributed among the queues. When each of queues contains k packets, the probability that the packet loss occurs is given by the following equation by assuming an exponentially distributed packet length [Cal00].

$$P_l(k) = e^{-m_k^{(i)}(B-1)D} \sum_{j=0}^{k-1} \frac{[m_k^{(i)}(B-1)D]^j}{j!}, \quad k > 0 \qquad (18)$$

We note here that in actual, we need to take account of the buffer occupancy (i.e., the unfinished work according to the terminology of the queueing theory), but we only consider the number of jobs queued in each buffer for simplicity. The newly arriving packet is lost if none of W queues can accept it. Thus, the state-dependent packet arrival rate for the next iteration is given by the following equations:

$$l_k^{(i+1)} = \begin{cases} l, & k < W \\ l\left\{1 - \left[P_l(\lfloor k / W \rfloor)\right]^W\right\}, & k \geq W \end{cases} \qquad (19)$$

That is, when the number of packets in the system exceeds W, the newly arriving packet is lost if buffer occupancies exceed the buffer size at all of queues.

Appendix B: Analysis for Optical Delay Line Buffers

We follow [Cal00] to analyze the optical buffering system with fiber delay lines. When all the servers are busy, the packet is stored at the shortest queue if the room is available. Since the granularity of the delay line buffer is D, we need an additional time given by Eq. (2), during which the server is idle. Since this extra delay is introduced when the arriving packet finds no idle servers, the mean of the fictitious packet length $1/\mu_e$ is approximately given by the following equation

$$1/m_e = p_0/m + (1-p_0)(1/m + D/2) \qquad (20)$$

where p_0 is the probability that all servers are idle, which is given by Eq. (18) of Appendix A. By using it for next iteration as a new value of $1/\mu$, we can determine the effect of the optical delay line buffers.

References

[Ban00] J. Bannister, J. Touch, A. Willner, and S. Suryaputra, "How Many Wavelengths Do We Really Need? A Study of the Performance Limits of Packet over Wavelengths," *Optical Networks Magazine*, pp.17-28, April 2000.

[Cal00] F. Callegati, "Optical Buffers for Variable Length Packets," *IEEE Communications Letters*, Vol.4, No.9, pp.292-294, September 2000.

[Chl96] I. Chlmtac, A. Fumagalli, L.G. Kazovsky, P. Melman, W.H. Nelson, P. Poggiolini, M. Cerisola, A.N.M.M. Choudhury, T.K. Fong, R.T. Hofmeister, C.–L. Lu, A. Mekkittikul, D.J.M. Sabido, IX, C.–J. Suh, and E.W.M. Wong, "CORD: Contention Resolution by Delay Line," *IEEE Journal on Selected Areas in Communications*, Vol.14, pp.1014-1029, 1996.

[Dan98] S.L. Danielson, P.B. Hansen, and K.E. Stubkjaer, "Wavelength Conversion in Optical Packet Switching," *Journal on Lightwave Technology*, Vol.16, pp. 2095-2108.

[Dan00] D.O. Awduche, Y. Rekhter, J. Drake, and R. Coltun, "Multi-Protocol Lambda Switching: Combining MPLS Traffic Engineering Control with Optical Crossconnects," *IETF Internet Draft*, draft-awduche-mpls-te-optical-02.txt.

[Dav98] B. Davie, P. Doolan, and Y. Rekhter, Switching in IP Networks - IP Switching, Tag Switching, and Related Technologies, Morgan Kaufmann, 1998.

[Fra99] A. Franzen, H. Sotobayashi, K. Kitayama, and I. Andonovic, "Demonstration of a High Resolution Synchronizer to Facilitate Payload Recovery at an Optical Node," *IEEE Photonic Technology Letters*, Vol.11, pp.1671-1673, 1999.

[Ge00] A. Ge, L. Tancevski, G. Castanon, L.S. Tamil, "WDM Fiber Delay Line Buffer Control for Optical Packet Switching," in *Proceedings of OptiComm 2000: Optical Networking and Communications*, pp.247-256, 2000.

[Hun98] D.K. Hunter, W.D. Cornwell, T.H. Gilfedder, A. Franzen, I. Andonovic, "SLOB: A Switch with Large Optical Buffers for Packet Switching," *Journal of Lightwave Technology*, Vol.16, No.10, pp.1725-1736, October 1998.

[Ito00] T. Ito, K. Fukuchi, K. Sekiya, D. Ogasahara, R. Ohhira, T. Ono, "6.4TB/s (160x40Gbit/s) WDM Transmission Experiment with 0.8bit/s/Hz Spectral Efficiency," *European Conference on Optical Communication (ECOC2000)*, PD-1.1, Munich, Sept. 2000.

[Kes98] S. Keshav and R. Sharma, "Issues and Trends in Router Design," *IEEE Communications Magazine*, pp.144-151, May 1998.

[Kit99] K. Kitayama and N. Wada, "Photonic IP Routing," *IEEE Photonic Technology Letters*, Vol.11, pp.1689-1691, 1999.

[Hal98] K.L. Hall and K.A. Rauschenbach, "All-Optical Buffering of 40-Gb/s Data Packets," *IEEE Photonic Technology Let-*

ters, Vol.10, pp.442-444, 1998.

[Lin96] H.-C. Lin and C.S. Raghavendra, "An Approximate Analysis of the Join the Shortest Queue (JSQ) Policy," *IEEE Transactions on Parallel and Distributed Systems*, Vol.7, No.3, pp.301-307, March 1996.

[Lin97] S. Lin and N. McKeown, "A Simulation Study of IP Switching," *Proceedings of ACM SIGCOMM '97*, pp.15-24, September 1997.

[Pru86] P. Prucnal, M. Santro, and T. Fan, "Spread Spectrum Fiber Optic Local Area Network using Optical Processing," *Journal on Lightwave Technology*, vol.4, pp.307-314, 1986.

[Sal89] J. Salehi, "Code Division Multiple Access Techniques in Optical Fiber Networks - Part I: Fundamental Principles," *IEEE Transactions on Communications*, Vol.37, pp.824-833, 1989.

[Sob98] H. Sotobayashi and K. Kitayama, "325nm Bandwidth Supercontinuum Generation at 10Gbit/s using Dispersion-Flattened and Non-Decreasing Normal Dispersion Fibre with Pulse Compression technique," *Electron Letters*, vol. 34, pp. 1336-1337, 1998.

[Tan99] L. Tancevski, A. Ge, G. Castanon, and L.S. Tamil, "A New Scheduling Algorithm for Asynchronous, Variable Length IP Traffic with Void Filling," in *Proceedings of OFC '99*, Paper ThM7, (San Diego), February 1999.

[Tan00] L. Tancevski, S. Yegnanarayanan, G. Castanon, L. Tamil, F. Masetti, and T. McDermott, "Optical Routing of Asynchronous, Variable Length Packets," *IEEE Journal on Selected Areas in Communications*, Vol. 18, No. 10, pp.2084-2093, October 2000.

[Tak00] H. Taara, T. Ohara, K. Mori, K. Sato, E. Yamada, T. Morioka, K. –I. Sato, K. Junguji, Y. Inoue, and T. Shibata, "Over 1000 Channel Optical Frequency Chain Generation from a Single Supercontinuum Source with 12.5GHz Channel Spacing for DWDM and Frequency Standards," *European Conference on Optical Communication (ECOC2000)*, PD-3.1, Munich, Sept. 2000.

[WAN97] "WAN Packet Size Distribution," available at http://www.nlanr.net/NA/Learn/packetsizes.html, June 1997.

[Wad99] N. Wada and K. Kitayama, "10Gb/s Optical Code Division Multiplexing using 8-chip Optical Bipolar Code and Coherent Detection," *Journal on Lightwave Technology*, Vol.17, pp.1758-1765, 1999.

[Wad00] N. Wada and K. Kitayama, "Photonic IP Routing Using Optical Codes: 10Gbit/s Optical Packet Transfer Experiment," *2000 Optical Fiber Conference (OFC2000)*, WM51 (Baltimore), 2000.