<table>
<thead>
<tr>
<th><strong>Title</strong></th>
<th>Parallel transfer optical packet switches</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Author(s)</strong></td>
<td>Li, CY; Wai, PKA; Li, VOK</td>
</tr>
<tr>
<td><strong>Citation</strong></td>
<td>Journal Of Lightwave Technology, 2009, v. 27 n. 12, p. 2159-2168</td>
</tr>
<tr>
<td><strong>Issued Date</strong></td>
<td>2009</td>
</tr>
<tr>
<td><strong>URL</strong></td>
<td><a href="http://hdl.handle.net/10722/58824">http://hdl.handle.net/10722/58824</a></td>
</tr>
<tr>
<td><strong>Rights</strong></td>
<td>Creative Commons: Attribution 3.0 Hong Kong License</td>
</tr>
</tbody>
</table>
Parallel Transfer Optical Packet Switches

C. Y. Li, Member, IEEE, Ping K. A. Wai, Senior Member, IEEE, and Victor O. K. Li, Fellow, IEEE

Abstract—For efficient utilization of bandwidth in optical packet switching, the guard time $T_g$ between packets should only be a small fraction of the packet transmission time $T_p$. Since the guard time $T_g$ of existing packet switching approaches must be larger than the reconfiguration time $T_{sw}$ of optical switches, this imposes a stringent demand on the switch reconfiguration time $T_{sw}$ as the transmission rate of optical fibers increases. By using batch transfer of packets or multiple switching fabrics in parallel, the requirement on the switch reconfiguration time can be significantly relaxed. The utilization of the transmission links can be greatly improved because the guard time between packets is no longer constrained by the switch reconfiguration time.

Index Terms—Blocking probability, optical switch, slotted optical network, switch reconfiguration time.

I. INTRODUCTION

O PTICAL network is one of the technologies that can provide the required transmission bandwidth for the rapidly growing communication traffic. Although terabits per second point-to-point transmission has been realized [1], a light path (wavelength channel) must be set up before any two nodes can exchange packets [2]. Owing to the lack of sophisticated optical signal processing devices and effective means to buffer light, all-optical packet switching is still in the research stage [3]. A more feasible approach of realizing optical packet switching is to optically switch the data packets but electronically process the packet headers for routing information. However, even such a hybrid approach is still difficult to realize.

One problem in implementing a practical optical packet-switched network is the difficulty of guaranteeing high-bandwidth utilization when the fiber transmission rate is high. In packet switched networks, a guard time $T_g$ between packets is required to prevent packets from interfering with each other. In existing packet switches, packets are switched/transferred one by one. The packet guard time $T_g$ must therefore be larger than the reconfiguration time $T_{sw}$ of the switches. Otherwise, accidental packet discard may occur. Recently, very fast all-optical switching has been demonstrated. Thus, the packet exchange rate and link utilization should only be limited by the processing speed of the packet headers [4]. However, fast optical switches with switching time $T_{sw}$ in nanosecond or picosecond ranges are only available in small sizes such as $2 \times 2$ [4]. Large optical switches with up to a thousand ports have also been demonstrated using the microelectromechanical system (MEMS) technology but the required switch reconfiguration time $T_{sw}$ is of the order of milliseconds [5]. Since no data transmission can occur during the guard time $T_g$, a large guard time $T_g$ will lead to low transmission bandwidth utilization. As the fiber transmission rate increases, the switch reconfiguration time $T_{sw}$ will become increasingly significant in the determination of the transmission bandwidth utilization.

In optical burst switching (OBS) networks [6], we have proposed to reduce the negative impact of large switch reconfiguration time $T_{sw}$ by reconfiguring the optical switches before the arrival of packets [7]. This approach, however, requires retrieval of prior information of the switch status from the OBS reservation signaling, and may not be directly applicable to other types of optical packet-switched networks such as the slotted optical networks. Many important networks are slotted, for example the asynchronous transfer mode (ATM) networks [8] and the deflection routed networks [9]. The size of a time slot is in general a compromise between different considerations of traffic and network performance. Consequently, increasing the packet size to maintain reasonable throughput is not feasible.

Although slotted networks in general do not provide prior information of the switch status as in OBS networks, the features of fixed packet size and synchronous transmission allow other ways to tackle the problem such as by using batch transfer of packets or multiple switching fabrics working in parallel. Both methods can significantly relax the constraint imposed on the packet guard time $T_g$ by the switching fabric reconfiguration time $T_{sw}$. In batch transfer, a packet is not immediately routed to its desired output when the packet arrives at the input of the switch. The optical switch waits for $K$ packet transmission times before transferring the packets to their desired outputs. Since the packets arrive continuously, the optical switch may have to process at most $K$ packets in a batch. The required inputs and outputs of the switching fabric used in the proposed switch architecture should therefore be $K$ times that of the switches if batch transfer is not used. Since the number of connecting links between nodes (switches) in the network remains unchanged, 1-to-$K$ packet serial-to-parallel and $K$-to-1 packet parallel-to-serial converters are used. As the proposed optical packet switch transfers $K$ packets in a batch from inputs to outputs during $K$ packet transmission times, the time available for the switch reconfiguration is roughly $K - 1$ packet transmission times.

Another way to relax the constraint on the packet guard time $T_g$ is to use $K$ switching fabrics in parallel such that each switching fabric transfers only one packet during each $K$
packet transmission time. Hence, each switching fabric will have $K - 1$ packet transmission times to reconfigure itself. If the $K$ switching fabrics are scheduled properly, the guard time $T_g$ between packets can also be smaller than the required time $T_{sw}$ for a single switching fabric reconfiguration. In this paper, we propose and analyze the switch architectures that can be used to implement batch packet transfer or multiple switching fabrics in parallel in order to relax the requirement on the switching fabric reconfiguration time $T_{sw}$. In Section II, we describe the proposed multisolot batch-transfer (MSBT) switch architecture. The operation of packet serial-to-parallel transmission conversion and the timing for packet transfer are discussed. In MSBT, the added packet delay is large and requires large switching fabric. In Section III, we discuss the multifabric sequential transfer (MFST) architecture which assumes multiple switching fabrics in parallel. To further reduce the added packet delay in the MFST switch, we decouple the routing information from the packets. We describe MFST with pilot message (MFST/PM) switch architecture in Section IV. Section V gives the performance evaluation of the proposed switches including the link utilization, added packet delay, packet loss performance, and delay variance. We briefly discuss the implementation consideration in Section VI. Finally, we conclude the paper in Section VII.

II. MULTISLOT BATCH-TRANSFER SWITCH

Fig. 1 shows the proposed MSBT switch architecture of a $2 \times 2$ optical switch where $I_1, I_2$ and $O_1, O_2$ are the input and output links, respectively. A $2 \times 2$ optical switch is used as an example for convenience of illustration. The number of switch inputs and outputs in practical applications such as wavelength division multiplexed (WDM) networks can be up to hundreds. In Fig. 1, $K$ duplicates of an optical signal from an input link $I_1$ are made by using the $1 \times K$ optical splitter. Each of the duplicated signals is delayed by $D_k$, $k = 1, \ldots, K$ with fiber delay lines (FDLs) and is sent to the inputs $I_{k,1}, k = 1, \ldots, K$ of the $2K \times 2K$ switching fabric. We assume that the switching fabric is a commercially available internally nonblocking optical switch. The optical splitter and the FDLs form a simple optical switch. The optical splitter and the FDLs form a simple optical switch.
I_1,3 of the switching fabric, respectively. After reading the address of packet 3 at time t_1, the MSBT switch starts to configure the switching fabric to prepare for packet transfer at time t_2 = t_0 + D_3 - T_{sw}. The packets 1, 2, and 3 are finally transferred to the switch outputs in time duration t_3 - t_4 after the completion of the switching fabric internal path setup. From Fig. 2, one observes that the switch can also start the switching fabric internal path configuration at time (t_0 + D_3 - T_{sw}) or (t_0 + D_1 - T_{sw}) instead with one or two empty output slots in the initial round of packet transfer. The average added packet delay, however, remains unchanged. In Fig. 2, the switch reads packet 6 for the next round of the packet transfer process at time t_5 and starts the switching fabric reconfiguration at time t_6. The switching fabric is therefore idle for a period of T_{idle} between the two reconfigurations. T_{idle} is smaller than a slot time T_{slot} and should be minimized for transmission bandwidth efficiency. We will discuss it in more detail when we derive the equation for the minimum value of K.

In Fig. 1, a K x 1 optical multiplexer (MUX) and the associated K sets of FDLs form an optical packet parallel-to-serial transmission converter. Packets on the outputs O_{i1} to O_{iK} of the switching fabric are individually delayed and sent to the K x 1 MUX to combine into the optical signal on output link O_k. As the delay added to each packet should be a constant, the delays of the FDLs on the outputs O_{i1} to O_{iK} are the complement of that on I_{i1} to I_{iK}, i.e., D_{i1} = D_{iK}. Also, a K x 1 optical coupler can be used instead of the K x 1 MUX in Fig. 1 if we assume that the switching fabric turns off its outputs during the reconfiguration. Otherwise, a fast K x 1 optical switch is required for implementing the MUX.

For the proposed MSBT switch to operate as shown in Fig. 2, we should appropriately choose the values of K and D_i, i = 1, ..., K. Since the switching fabric can transfer the input packets only after it completes the reconfiguration process, all packets must be delayed by at least T_{cp} + T_{sw}. In principle, any value larger than T_{cp} + T_{sw} can be used for D_i, e.g., D_1 is set to 1.1 times T_{cp} + T_{sw} in Fig. 2. To minimize the added packet delay, however, we should choose

$$D_1 = T_{cp} + T_{sw}$$

Owing to the requirement of packet serial-to-parallel transmission conversion and minimizing the added packet delay, the delay difference between two adjacent FDLs should be equal to a slot time such that

$$D_k = D_{k-1} + T_{skt}, \quad 2 \leq k \leq K.$$  \hspace{1cm} (2)

Equation (2) and Fig. 1 assume that the output lookup time T_{cp} of a packet is smaller than a slot time T_{skt}, or multiple processors are used such that outputs of different packets are looked up in parallel. Otherwise, the difference between D_k and D_{k-1} may be larger than a slot time T_{skt}.

To compute the minimum value of K, we further assume that the MSBT switch can read input packets to prepare the next round packet transfers independently of the current status of the switching fabric. The MSBT switch can also schedule multiple reconfigurations for the switching fabric. Consequently, the value of T_{cp} will have no effect on the minimum value of K and the channel utilization. Since the switch transfers K packets from an input link I_k after each switching fabric reconfiguration, the time between two switching fabric reconfigurations is therefore K T_{skt}. All transmissions between the inputs and outputs of the switching fabric must be completed before the next switching fabric reconfiguration. Since T_{skt} = T_d + T_{tg}, the maximum value of the switching fabric reconfiguration time (T_{sw})_{max} becomes (K - 1) T_d + K T_{tg} if the T_{skt} in Fig. 2 is zero. Hence, the required minimum value of K can be written as

$$K_{min} = \left[\frac{(T_{sw} + T_d)}{(T_{tg} + T_d)}\right]$$  \hspace{1cm} (3)

where [x] is the smallest integer larger than x. In Fig. 2, T_{sw} is 1.7 time slots and the required K_{min} is therefore equal to 3. From (3), K_{min} is equal to one only if T_{sw} \leq T_{tg}, i.e., the switching fabric is fast enough. Otherwise, we have to extend T_{tg} to prevent collisions between packets as in the traditional approaches.

As the transmission rate of optical fiber grows, the reconfiguration time of large optical switches will likely be comparable to the packet transmission time in the near future. As shown in Section V, the proposed optical switch architecture will provide an excellent way to increase bandwidth efficiency. High bandwidth utilization can be achieved at the expense of added delay D_k per intermediate node. The added delay will become trivial if a packet transmission time is much smaller than the end-to-end propagation delay. However, the requirement of an NK x NK switching fabric inside an N x N switch will become a problem if N is large. Traditionally, one can break a large switch into smaller switches arranged in a multistage architecture to save the hardware cost. For example, a three-stage Clos switch architecture with first and last stages of N sets of K x K switches, and a second stage of K sets of N x N switches can replace the NK x NK switching fabric without any difference in blocking performance [12]. In the MSBT switch application, we may only use a two-stage switch architecture instead to further reduce the required hardware. However, for optical signal quality considerations, it may be better to keep the switching fabric to one stage [10].

### III. MULTIFABRIC SEQUENTIAL TRANSFER SWITCH

Another viable solution is to replace the NK x NK switching fabric with K sets of N x N switching fabrics in parallel if we transfer a packet only to its original time slot at the output link. Fig. 3 is a proposed multifabric sequential transfer (MFST) switch architecture for a 2 x 2 optical switch. Again, in practical applications, the MFST switch architecture will be an N x N optical switch where N can be hundreds or more. With K sets of 2 x 2 switching fabrics being connected in parallel as shown in Fig. 3, each packet can only be transferred to its original time slot at the output link. Similar to MSBT, K duplicates of an optical signal from an input link I_k are made by using the I x K optical splitter in Fig. 3, and each of them is sent to the input
required for packet output lookup and switching fabric reconfiguration is also equal to 1.7 time slots, i.e., $T_{cp} + T_{sw} = 1.7T_{slot}$. We also assume that all switching fabrics detect the packet addresses at the input link $I_1$. However, the switching fabrics are scheduled to operate in sequence such that switch fabric $k$ only starts its packet output lookup and reconfiguration at time slots $3Z + k$, where $Z$ is a nonnegative integer and $k = 1, \ldots, 3$. For example, switching fabric 1 only takes care of the packets in time slots 1, 4, 7, $\ldots$ as shown in the Fig. 4. We assume that the packet transfer delay from the inputs to the outputs of a switching fabric is negligible. In Fig. 4, a packet 1$^*$ (it may come from inputs $I_{1,4}$ or $I_{2,1}$) is sent to output $O_{1,1}$ during time period $t_1$ to $t_2$. The switching fabric 1 then waits until time $t_3$ and takes $T_{cp} + T_{sw}$ time for the packet output lookup and internal path reconfiguration to transfer packet 4$^*$ to output $O_{1,1}$ at time $t_5$. During the time period $t_2$ to $t_5$, switching fabrics 2 and 3 process the input packets 2 and 3 and transfer packets 2$^*$ and 3$^*$ to outputs $O_{1,2}$ and $O_{1,3}$ in sequence. Similarly, switching fabrics 3 and 1 will transfer packets 3$^*$ and 4$^*$ to outputs $O_{1,3}$ and $O_{1,1}$ during the reconfiguration of switching fabric 2. As the switching fabrics shift their operations in sequence, the proposed MFST switch in Fig. 3 can transfer packets between its inputs and outputs without any interruption.

In Fig. 4, the delay $D_0$ can be set to any value larger than $T_{cp} + T_{sw}$ similar to that of $D_1$ in Fig. 2. The required number of switching fabrics can also be computed from (3). Apart from replacing the $NK \times NK$ switching fabric with $K$ smaller $N \times N$ ones, one advantage of the MFST switch architecture shown in Fig. 3 is that $D_0$ is also the total per node added delay while that of the MSBT in Fig. 1 is that $D_K = D_1 + (K - 1)T_{slot}$ from (2). With the assumption $D_0 = D_1$, the reduction of the additional delay is $(K - 1)T_{slot}$ per intermediate node, where $K \geq 2$. This will be useful if the end-to-end propagation delay is not much larger than $T_{slot}$ and the application is delay sensitive. However, $D_0$ must be larger than $T_{cp} + T_{sw}$ which can cause the per node added delay of the MFST switch to be larger than $T_{slot}$. To further reduce the added delay, we need to decouple the packet address information from the packets so that a node switch can have prior information to configure the switching fabrics before the arrival of the input packets.

IV. MULTIFABRIC SEQUENTIAL TRANSFER WITH PILOT MESSAGE

Fig. 5 is the proposed switch architecture for multifabric sequential transfer with pilot message (MFST/PM) of a $2 \times 2$ optical switch. It is similar to that of the MFST in Fig. 3 apart from an additional switching control processor (SWCP). The switching fabrics no longer read the packet address information from the input link $I_i$. SWCP reads the packet address information from the control channels $C_{in\rightarrow 1}$ and $C_{in\rightarrow 2}$. We assume that the control channel $C_{in\rightarrow i}$ carries messages of packet address information $T_{dif\rightarrow i}$ time ahead of the packets on input link $I_i$, where $i = 1, 2$. We also assume that SWCP needs time $T_{cp}$ to detect and complete the lookup for each packet address. Once the packet output lookup is completed, SWCP instructs the switching fabrics to reconfigure the internal paths. At the same time, SWCP sends pilot messages to control channels $C_{out\rightarrow 1}$ and $C_{out\rightarrow 2}$ to inform the subsequent nodes about the
Fig. 5. Proposed multifabric sequential transfer with pilot message (MFST/PM) switch architecture for a $2 \times 2$ optical switch. It is similar to that of the MFST in Fig. 3 apart from an additional switching control processor (SWCP).

Fig. 6. Timing diagram for the packets at input link $I_1$, at inputs $I_{1,1}$ to $I_{1,3}$, outputs $O_{1,1}$ of the switching fabric, and the pilot messages at control channels $C_{in-1}$ and $C_{out-1}$ of the proposed MFST/PM $2 \times 2$ switch in Fig. 5 with $K = 3$.

Fig. 7. Maximum link utilization of switch outputs for the proposed MSBT, MFST, and MFST/PM switch architectures. The curves with crosses, asterisks, circles, and squares are the maximum output link utilization of the proposed switches with $K = 2, 3, 4,$ and $5$, respectively. The curve with pluses is that of the normal optical switch, i.e., the proposed switches with $K = 1$.

packets that will be sent to outputs $O_1$ and $O_2$ with an offset time $T_{offset}$ later. Hence, the subsequent nodes can also arrange the switching fabric settings similarly at $T_{offset}$ time before the arrival of packets. The detail of the operation is illustrated using the example in Fig. 6.

Fig. 6 is the timing diagram for the packets at input link $I_1$, at inputs $I_{1,1}$ to $I_{1,3}$, outputs $O_{1,1}$ of the switching fabric, and the pilot messages at control channels $C_{in-1}$ and $C_{out-1}$ of the proposed MFST/PM $2 \times 2$ switch with $K = 3$. We assume the same $T_{CP}$ and $T_{SW}$ as that of Figs. 2 and 4. In Fig. 6, the pilot message 1 arrives at time $t_0$. SWCP takes time $T_{CP}$ to look up the output and therefore the switching fabric starts the internal path reconfiguration at time $t_1$. To compensate for the pilot message processing time at SWCP, the FDLs at each input link $I_i$ has delay value of $D \geq T_{CP}$. Therefore, at time $t_1$, SWCP has the complete information of the whole switch in the next $T_{offset}$ time, e.g., packet 1* will be sent to output $O_{1,1}$ at time $t_2$. If packet 1* is a new packet, we assume that it is delayed at least $T_{CP} + T_{SW}$ time before it is allowed to enter the switch, i.e., packet 1* has to arrive at the switch before time $t_0$ such that SWCP will include it in the output assignment at $t_1$. Hence, SWCP sends out the pilot message 1 to the control channel $C_{out-1}$ at time $t_1$ without waiting for the completion of the switching fabric reconfiguration. With the assumption of using the same routing paths of their associated packets, pilot messages will arrive at the subsequent nodes $T_{offset}$ time ahead, and the subsequent nodes can preconfigure their switching fabrics accordingly.

The minimum added delay of the MFST/PM switch in Fig. 5 will be $H \times T_{CP} + T_{SW}$ for an $H$ hop path if we assume that $D = T_{CP} + T_{offset} = T_{SW}$, and a packet is filtered out before the input of the destination switch. The minimum delay includes the $T_{CP} + T_{SW}$ initial delay at the source and the $T_{CP}$ delay at each intermediate node. The intermediate node switching fabric reconfiguration time, however, becomes irrelevant to the end-to-end delay because it overlaps with the delay time at the source and the node-to-node traveling time of the packets, just as in OBS [6], [7]. Since the $H$ hop path minimum added delays of the MSBT switch in Fig. 1 and MFST switch in Fig. 3 are $H \times (T_{CP} + T_{SW} + (K - 1)T_{slot})$ and $H \times (T_{CP} + T_{SW})$, respectively, the delay savings with the MFST/PM switch in Fig. 5 will be significant if the reconfiguration time of the switching fabric is large.

V. PERFORMANCE EVALUATION

A. Link Utilization

Fig. 7 shows the maximum (achievable) link utilization of the switch outputs for the proposed MSBT, MFST, and MFST/PM switch architectures provided that the guard time for preventing the crosstalk between packets is not required. Recall that the utilizations of the different proposed switches are the same. In Fig. 7, the curves with crosses, asterisks, circles, and squares are the maximum output link utilizations of the proposed switches with $K = 2, 3, 4,$ and $5$, respectively,
where, $K$ is the packet serial-to-parallel conversion ratio of the MSBT switch and the number of switching fabrics of the MFST and MFST/PM switches. The curve with pluses is that of the normal optical switch, i.e., it is equal to the proposed switches with $K = 1$. The horizontal axis is the switch reconfiguration time normalized by the packet transmission time, i.e., $T_{sw}/T_d$. The maximum link utilization in the vertical axis is calculated using $T_d/(T_d + T_g)$ where the interpacket guard time $T_g$ is set to the minimum value to get the required $K$ in (3) with the given $T_{sw}$ and $T_d$. From (3), we have

$$T_g = \begin{cases} \left(\frac{T_{sw}}{T_d} - K + 1\right)/T_d & \text{if } \left(\frac{T_{sw}}{T_d}\right) > K - 1 \\ 0 & \text{otherwise.} \end{cases}$$

We assume that the delays $D_1$ to $D_K$ of the FDLs in Fig. 1 are changed accordingly. In Fig. 7, we vary the normalized switching time $T_{sw}/T_d$ between 0.01 and 10 to show the effect of switch reconfiguration time variations. For a normal optical switch, the switching overhead in general should only be a small fraction of the packet length, e.g., $T_{sw}/T_d < 0.1$. Otherwise, the link utilization will drop rapidly as shown in Fig. 7. The maximum link utilization for normal optical switch drops to 0.5 when $T_{sw}/T_d = 1$. In contrast, all curves for the proposed switches remain at unity until $T_{sw}/T_d$ is larger than $K - 1$. With the proposed switches, the tolerable switch reconfiguration time $T_{sw}$ increases from a fraction of a packet transmission time to several packet transmission times. While packets typically have lengths of kilobytes, we can use the switching fabrics with reconfiguration time a thousand times of that normally required for optical packet switching, e.g., the maximum length of an Ethernet packet is 1522 bytes [13].

**B. Added Packet Delay**

Fig. 8 shows the average added end-to-end delay to the packets in an $8 \times 8$  Manhattan Street Network (MSN) [14] with the proposed switches that we have presented in Sections II, III, and IV. The curves with dots, triangles and diamonds are the average end-to-end added delay of the $8 \times 8$ MSN network with MSBT, MFST, and MFST/PM switches, respectively.

At $T_{sw} = 10$, the added delays of the MSBT, MFST, and MFST/PM switches are 91.2 (not shown in Fig. 8), 46.1, and 9.5 time slots ($T_{sw}/T_d$), respectively. As shown in the Fig. 8, the MFST/PM switch can significantly reduce the added delay if the switch reconfiguration time $T_{sw}$ is a multiple of the packet transmission time $T_d$. Since the MFST/PM switch requires additional signaling of pilot messages in the network, the MFST approach in general should be used unless $T_{sw}$ is really large.

**C. Packet Discarding Performance**

All the proposed switch architectures for MSBT, MFST, and MFST/PM are internally nonblocking switches. However, the MSBT switch can further reduce output blocking by allowing the packets to shift from their original time slots at the output of the switch. This function can be implemented without time slot interchanger by transferring a packet from input $I_{i,j}$ (a packet from the slot $j$ of the $K$ packet batch on input link $i$) of the switching fabric to any output $O_{m,n}$ for $1 \leq n \leq K$, not just to output $O_{m,j}$ only, where $m$ is the preferred output of the packet. Thus, $K$ time slots are available to the packet instead of only the original time slot. To demonstrate the effect of $K$ on reducing the packet loss probability of the MSBT switches, we define $p_f(L)$ as the probability of having $L$ other contending packets for an output link $O_k$ when a packet is at input $I_{i,k}$, and the MSBT switch is starting to look up the outputs of the input packets, i.e., at time $t_2$ in Fig. 2, where $1 \leq i \leq N$, $1 \leq k \leq K$, $0 \leq L \leq (NK - 1)$, and $N$ is the number of input/output links. We assume that the arriving packets choose the output links at random. Hence, we have

$$p_f(L) = \sum_{M=L}^{NK-1} q(M) \left(\frac{M-1}{N}\right)^L \left(\frac{N-1}{N}\right)^{M-L}$$

(4)
where \( q(M) \) is the probability of \( M \) other packets arriving at other inputs of the switching fabric. \( p_i(L) \) is the binomial distribution with parameters \( M, L, \) and \( 1/N \) when \( M \) is fixed. Assuming that the packets arrive uniformly at all input links, \( M \) is also a binomial random variable. We have

\[
q(M) = \binom{NK-1}{M} \rho^M (1-\rho)^{NK-M-1}
\]

(5)

where \( \rho \) is the utilization of the input links. Since the packets choose output links at random, the average packet loss probability is equal to that of the packet being discarded at any output link \( Q_i \). Assuming that each packet has the same priority, the average packet loss probability \( B \) can be written as

\[
B = \sum_{L=K}^{NK-1} p_i(L) \frac{L-K+1}{L+1}.
\]

(6)

Fig. 9 shows the packet loss probability of the MSBT switch with two input and two output links (\( N = 2 \)) for different values of \( K \). As in Fig. 7, the curves with crosses, asterisks, circles, and squares represent the loss probabilities of the proposed MSBT switch with \( K = 2, 3, 4, \) and 5, respectively. The curve with pluses is the packet loss probability of the normal optical switch. It is also the loss probability of the MSBT switch with \( K = 1 \) and that of the MFST and MFST/PM switches. From Fig. 9, the packet loss probability of the MSBT switch decreases when \( K \) increases, especially when the system is lightly loaded. Thus, the proposed MSBT switch architecture can greatly improve packet loss performance in addition to easing the constraint on the switch response time. However, the reduction in loss probability (compared with the loss probability when \( K = 1 \)) decreases when the link utilization increases.

D. Delay Variance

As we have discussed in Section II, the FDLs in the outputs and inputs of the MSBT switch are complementary to each other, i.e., \( D_{iK} + D_{iK} = k \) is a constant. Transferring a packet to a time slot other than its original time slot at the output link will cause a delay fluctuation of up to \( \pm (K - 1) \) time slots. Using (6), we define \( B_N(K) \) as the packet loss probability on an \( N \times N \) MSBT switch with \( K \) serial-to-parallel packet conversion ratio. Hence, \( B_N(1) \) is also the loss probability of the MSBT switch if only the original time slot can be used for the packet transfer, and \( P_n = (B_N(1) - B_N(K))/(1 - B_N(K)) \) is the probability of a packet not transferred to its own time slot. We assume that a packet is assigned to any available one of the \( K \) time slots at random if the default time slot is unavailable. Let \( x \) be the delay fluctuation of a transferred packet. If a packet arrives at the \( k \)th of the \( K \) time slots, \( x \) will have probability \( (1-P_n) \) to be 0 and \( P_n/(K - 1) \) to be one of the \( K-1 \) values of \((-k-1), \ldots, -1, 1, \ldots, K-k\), where \( 1 \leq k \leq K \). Since packets arrive randomly at each time slot and switch input, the probability distribution of \( x \) is \( P(x = z) = P_n \times (K - |z|)/(K - 1) \), where \( |z| \) is the absolute value of \( z \), and \(- (K - 1) \leq z \leq (K - 1) \). We have the expectation \( E(x) = 0 \) and the variance \( \text{Var}(x) = E(x^2) - [E(x)]^2 \) to be the delay variance of a transferred packet in MSBT switches.

Fig. 10 shows the packet delay variance of the MSBT switch with \( N = 2 \) and different values of \( K \). In Fig. 10, there is no packet delay variance for other proposed switches, i.e., the curve with the pluses. For the MSBT switch, the delay variance increases with \( K \) and also increases with the link utilization in the range 0–0.8. Figs. 9 and 10 show that the packet loss probability of the MSBT switches may be reduced at the expense of additional delay and delay variance to the packets. We can maintain the packet sequence integrity by optimizing the switch output port to packet assignment, but the packet delay variance cannot be avoided unless additional optical hardware such as
optical time slot interchangers are used. Since packets can accumulate significant delay variance along the path, large buffers are required at the end nodes to smooth out the jitters if real time traffic is carried in the networks.

VI. IMPLEMENTATION CONSIDERATIONS

In this paper, we propose a way to increase the throughput of optical networks by using batch transfer of packets or multiple switching fabrics in parallel. We have proposed the MSBT, MFST, and MFST/PM switch architectures to overcome the stringent demand on the optical switching speed and the relative duration of the guard time between packets. Many issues must be considered for the realization of the proposed switches, including the power consumption and dissipation, scalability, length of the optical delay lines required, complexity of the scheduler, and cascadability. In the following, we will discuss some of the implementation issues.

Power consumption is an important issue in the design of future routers because power consumption by data centers nowadays can rival that of a small town. For the proposed switches, power consumption in the worst case is proportional to the square of the number of ports $N$ if technologies such as the semiconductor optical amplifier (SOA) crossbar structure are used to build the switching fabrics [4]. Since MBST switches use a single $KN \times KN$ switching fabric, the worst case power consumption will be proportional to $K^2 N^2$ times the power consumed by individual switching elements. Large power consumption will also affect the scalability of multiple routers in parallel configurations. The proposed MFST and MFST/PM switches use $K$ switching fabrics in parallel and therefore the power consumption is proportional to $K$ times that of individual switching fabric. In general, the MFST and MFST/PM switches have smaller power consumption, and will have better scalability than MSBT switches.

It is well known that FDLs are large and bulky. One meter of FDLs introduces 5 ns of delay. A packet with transmission time $T_{d1}$ of 1.5 $\mu$m needs at least 300 m of FDLs to store. In the proposed MSBT switches, delays of multiple packet time duration are required for the packet serial-to-parallel and parallel-to-serial transmission converters. Thus, kilometer long FDLs will have to be used. Hence, MSBT switches will face the same problems as other switches with optical buffers if the number of ports $N$ and the serial-to-parallel ratio $K$ are large. Since MFST and MFST/PM switches require smaller packet delay (it is independent of $T_{d1}$) and only one set of FDLs is used for an input port, these switches will use much less FDLs than MSBT switches. From this consideration, the smaller packet delay of MFST and MFST/PM switches not only can improve the system performance but also has advantages in implementation of the proposed switches.

Using a 1-to-$K$ optical power splitter ($K$-to-1 power combiner) to implement the packet de-multiplexer (packet multiplexer) can introduce a large insertion loss. For example, in a 1-to-64 splitter, the packets will experience more than 36 dB of power losses. This is another reason that we prefer to have a small $K$. As shown in Section V-A, we can use switching fabrics with reconfiguration time $T_{sw}$ up to $(K - 1)$ packet transmission time $T_{d}$ in our proposed switches to have nearly 100% link utilization. If $T_{d}$ is in the order of microseconds, optical switches with microsecond reconfiguration time can be used for the switching fabrics. Thus, MEMS switches are in general not suitable as the switching fabrics because of their millisecond reconfiguration time. Recently, an 8 $\times$ 8 PLZT (Pb$_x$La$_{1-x}$)(Zr$_y$Ti$_1-y$O$_3$) electro-optic switch (port count extension capability up to 64 ports) with a microsecond reconfiguration time has been demonstrated [15]. A 64 $\times$ 64 GaAs phased array electro-optic switches can even achieve a reconfiguration time of 30 ns but with the drawback of polarization dependence [16]. Hence, $K = 2$ or $K = 4$ should be generally sufficient. For a larger but moderate value of $K$, such as $K = 8$, we have proposed to use 1 $\times$ $K$ and $K$ $\times$ 1 optical switches to replace the splitters and combiners in all MSBT, MFST, and MFST/PM switches. Optical switches such as the 1 $\times$ 8 PLZT optical switches with potential loss of 6 dB and $T_{sw}$ around 10 ns are available [17]. If $T_{d}$ is so small that $K$ has to be large, optical power splitters and combiners will have to be used for packet de-multiplexing and multiplexing but then additional signal amplification is needed to compensate for the power losses. Of course, the system complexity will increase.

The proposed switches continually process the packets and the switch schedulers have to assign each packet a suitable output within a slot time, i.e., $T_{op} < T_{slot}$. When the optical fiber transmission rate increases and $T_{slot}$ becomes small, the performance of the switch scheduler will become a limiting factor for the system throughput. Many approaches have been used to improve the scheduler performance. They include efficient lookup algorithms [18], better routing table designs [19], hardware-based lookup methods [20], and elimination of the lookup process by using self-routing schemes [21]. In this paper, we focus on the transmission bandwidth overhead caused by the switch reconfiguration time, and simply assume that the switch schedulers have the required performance.

A physical control channel is not necessary for MFST/PM switches because we can embed the pilot messages into the earlier arriving packets though it may add a delay slightly larger than the minimum value, e.g., pilot messages 3 and 4 can be carried in packets 1 and 2 of Fig. 6. Since MFST/PM switches use the pilot messages to shorten the packet delay, it is necessary that the same type of switching fabrics (with the same reconfiguration time) is used at all nodes. In contrast, both MSBT and MFST switches do not need pilot messages and they have greater flexibility of using different types of switching fabrics at different nodes.

In this paper, we have only discussed the basic architectures of the proposed switches. Many services can be provided by the proposed switches with no or minor modifications. For example, using multichannel deflection routing to resolve packet contention has been proposed for networks with nodes of MSBT switches [22]. The analytical model for throughput delay performance of the MSBT networks has been derived. Also, multicast service can be implemented on MSBT switches if multiple delay switchable FDLs are used in the serial-to-parallel packet converter. Finally, the MSBT switches with switchable FDLs can provide other important services such as packet buffering and priority routing.
VII. CONCLUSION

We have studied the use of batch transfer of packets and multiple switching fabrics in parallel to relax the stringent constraint imposed by the switch reconfiguration time on optical packet switched networks. All three proposed switch architectures can provide the same bandwidth utilization improvement but with different added packet delays and implementation requirements.

In the MSBT switch architecture, we use packet serial-to-parallel transmission conversion to retrieve the future information of a batch of arriving packets and use the gained time to pre-configure the switching fabric. Although the MSBT switch can relax the requirement of switch reconfiguration time from a small fraction of the packet transmission time to multiple packet transmission times, it requires a large single switching fabric and may be difficult to implement. We therefore, also propose the multifabric sequential transfer (MFST) architecture that uses multiple smaller size switching fabrics instead. Apart from simplifying the implementation, the MFST also greatly reduces the added delay to packets compared to that of the MSBT switch. To further reduce the added packet delay, we decouple the routing information from the packets and propose the MFST with pilot message (MFST/PM) switch architecture.

Although the MSBT architecture uses a single large switching fabric and has the largest added packet delay, it can reduce the packet loss probability without extra hardware by not limiting packet transfer only to its original time slot at the output link. However, large buffers will be required at the end nodes of the network because of the increased delay variance.

REFERENCES

Victor O. K. Li (S’80–M’81–SM’86–F’92) received SB, SM, EE and ScD degrees in electrical engineering and computer science from Massachusetts Institute of Technology, Cambridge, MA, in 1977, 1979, 1980, and 1981, respectively. In 1981, he joined the University of Southern California (USC), Los Angeles, CA, and became Professor of Electrical Engineering and Director of the USC Communication Sciences Institute. Since September 1997 he has been with The University of Hong Kong, Hong Kong, where he is Associate Dean (Research) of Engineering, and Chair Professor of Information Engineering, Department of Electrical and Electronic Engineering. He was the Managing Director of Versitech Ltd., the technology transfer and commercial arm of the University, and on various corporate boards. His current research interests include information technology, including all-optical networks, wireless networks, and Internet technologies and applications.

Prof. Li was the Chair of the Computer Communications Technical Committee of the IEEE Communications Society during 1987–1989, and the Los Angeles Chapter of the IEEE Information Theory Group during 1983–1985. He co-founded the International Conference on Computer Communications and Networks (IC3N), and was the Chair of its Steering Committee during 1992–1997. He was also the Chair of various international workshops and conferences, including, most recently, IEEE INFOCOM 2004 and IEEE HPSR 2005. Prof. Li has served as an editor of IEEE Network, IEEE JSAC Wireless Communications Series, IEEE Communications Surveys and Tutorials, and Telecommunication Systems. He also guest edited special issues of IEEE JSAC, Computer Networks and ISDN Systems, and KICS/IEEE Journal of Communications and Networking. He is now serving as an editor of ACM/Springer Wireless Networks. Prof. Li has been appointed to the Hong Kong Information Infrastructure Advisory Committee by the Chief Executive of the Hong Kong Special Administrative Region (HKSAR). He was a part-time member of the Central Policy Unit of the Hong Kong Government. He was also with the Innovation and Technology Fund (Electronics) Vetting Committee, the Small Entrepreneur Research Assistance Programme Committee, and the Engineering Panel of the Research Grants Council. He was a Distinguished Lecturer at the University of California at San Diego, at the National Science Council of Taiwan, and at the California Polytechnic Institute. He has also delivered keynote speeches at many international conferences.

Prof. Li has received numerous awards, including, most recently, the PRC Ministry of Education Changjiang Chair Professorship at Tsinghua University, the UK Royal Academy of Engineering Senior Visiting Fellowship in Communications, the Outstanding Researcher Award of The University of Hong Kong, the Croucher Foundation Senior Research Fellowship, and the Order of the Bronze Bauhinia Star, Government of HKSAR, China. He is a Fellow of the Hong Kong Institution of Engineers and the IAE.