On Efficient Constructions of Optical Priority Queues

Jay Cheng, *Senior Member, IEEE*, Sheng-Hua Yang, Chun-Yung Wang, Hao-Hsuan Tang, and Bin Tang, *Member, IEEE*

Abstract

The design of optical buffers for packet contention resolution has been recognized as a key issue in all-optical packet switching. One of the most general buffering schemes is priority queues, which includes first-in first-out (FIFO) queues and last-in first-out (LIFO) queues as special cases. In a priority queue, each packet is associated with a unique priority upon its arrival, the packet with the *highest* priority is sent out from the queue whenever there is a departure request and there are packets in the queue, and the packet with the *lowest* priority is dumped from the queue whenever there is a buffer overflow. In this paper, we consider the constructions of optical priority queues by using a feedback system consisting of an optical (bufferless) crossbar switch and multiple optical FIFO multiplexers with delay one (FM1's) in the feedback path for buffering packets and feeding packets back to the switch. Such a feedback system is a generalization of that used in one of the authors' earlier attempt for the constructions of optical priority queues in [19]. We fix the *no-buffering* problem in [19] by using optical FM1's to replace the optical FIFO multiplexers (FM's) in [19], which enables us to successfully achieve an exact emulation of a priority queue. We improve the utilization of buffering capacity over that in [19] by routing packets to the optical FM1's according to their *buffering tags* instead of their *tags* as used in [19]. We also extend and generalize the construction in [19] and obtain a much larger class of constructions of optical priority queues. Our constructions are made possible by showing that the highest-priority (resp., lowest-priority) packet is always available at the input links of the switch whenever it needs to be routed to the departure (resp., loss) link, and by showing that there is no collision and there is no buffer overflow at any FM1 at any time so that there is no internal packet loss at any time. Our complexity analysis shows that by using a feedback system consisting of an optical $(M+2)\times(M+2)$ (bufferless) crossbar switch and M fiber delay lines, we can achieve a buffer size of $2^{O(\sqrt{\alpha M})}$, where α is a constant that depends on the parameters used in our constructions. Furthermore, we show that the best buffer size that we can achieve is $2^{O(\sqrt{4M/15})}$. Our result (exponential in $\sqrt{M}$) substantially improves on the best known result (polynomial in M) in the literature. Our numerical results show that the construction complexity of our constructions is lower than that of the construction in [19], and the actual saving, in terms of the number of 2×2 switches needed, by our constructions could be quite significant even in the tiny-buffer and small-buffer regimes.

Index Terms

FIFO multiplexers, optical buffers, optical queues, optical switches, priority queues.

The work of J. Cheng, S.-H. Yang, C.-Y. Wang, and H.-H. Tang was supported by the Ministry of Science and Technology of Taiwan under Grant MOST-105-2221-E-007-035-MY3 and Grant MOST-108-2221-E-007-017-MY2. This paper is an improved version of our work presented in part at the International Conference on Systems and Informatics (ICSAI'19), Shanghai, China, November 2–4, 2019, and in part at the IEEE International Conference on Computer and Communications (ICCC'19), Chengdu, China, December 6–9, 2019. To facilitate the reader's understanding of the routing policy in this paper, we have prepared slides that are available at https://www.ee.nthu.edu.tw/jcheng/publications/PQ-efficient-constructions.pptx

J. Cheng, S.-H. Yang, C.-Y. Wang, and H.-H. Tang are with the Department of Electrical Engineering, National Tsing Hua University, Hsinchu, Taiwan, and B. Tang is with the School of Computer and Information, Hohai University, Nanjing 211100, China (e-mail: jcheng@ee.nthu.edu.tw; s109061851@m109.nthu.edu.tw; s102032033@m102.nthu.edu.tw; s50405eric@gmail.com; cstb@hhu.edu.cn).

I. INTRODUCTION

Current packet-switched networks suffer from the notorious optical-electrical-optical (O-E-O) conversion, which is quite expensive and time-consuming, and hence cannot keep up with the pace of the growing optical fiber link capacity. A natural and attractive solution for overcoming the existing O-E-O hurdle and making good use of the tremendous bandwidth offered by optical fiber links is all-optical packet switching. However, optical random-access memory (RAM) is not available yet for contention resolution among packets competing for the same resources. Fortunately, we only need buffering schemes that can exactly emulate certain special arrival/departure patterns in many packet-switched networks. Such buffering schemes with special arrival/departure patterns are generally known as *queues* in the context of queueing theory.

In all-optical packet switching, the design of optical queues has been well recognized as a very challenging problem. In the last two-plus decades, there have been extensive studies on the constructions of a variety of optical queues by using fiber delay lines as the storage media and using optical (bufferless) crossbar switches to direct optical packets through the fiber delay lines in a carefully designed manner so as to achieve exact emulations of the optical queues. Such Switched-Delay-Lines (SDL) constructions of optical queues by using optical crossbar switches and fiber delay lines include output-buffered switches, first-in first-out (FIFO) multiplexers [1]–[12], FIFO queues, last-in first-out (LIFO) queues, priority queues [13]–[19], time slot interchanges, linear compressors, linear decompressors, non-overtaking delay lines, flexible delay lines, FIFO contractors, LIFO contractors, and absolute contractors. Due to space constraint, we only list the references [1]–[19] on optical FIFO multiplexers (FM's) and optical priority queues that are directly related to the constructions in this paper, and results on the other types of optical queues and results on fundamental complexity, performance analysis, and review articles for SDL constructions of optical queues can be found in the references therein.

The main research issue in SDL constructions of optical queues is on the design of the delays of the fiber delay lines and the design of the routing policy performed by the optical crossbar switches, which are closely related and highly coupled. As in most works on SDL constructions of optical queues in the literature, in this paper we consider the following discrete-time settings: (i) Time is slotted and synchronized. (ii) Packets are of the same size so that a packet can be transmitted through a link within a time slot. (iii) An optical $M \times M$ (bufferless) crossbar switch is a network element with M input links and M output links that can realize all of the $M!$ permutations between its inputs and its outputs. (iv) A fiber delay line with delay d is a network element with one input link and one output link that requires d time slots for a packet to traverse through. We note that variable-size packets can be easily taken care of by introducing packet segmentation at the source and packet reassembly at the destination. For reason of conciseness, in the rest of this paper we simply refer to time slot t as "slot t."

In this paper, we consider SDL constructions of optical priority queues. A priority queue with buffer size B has one arrival link, one departure link, one loss link, and one control input (see Figure 1(a)). Each packet is associated with a unique priority upon its arrival so that every

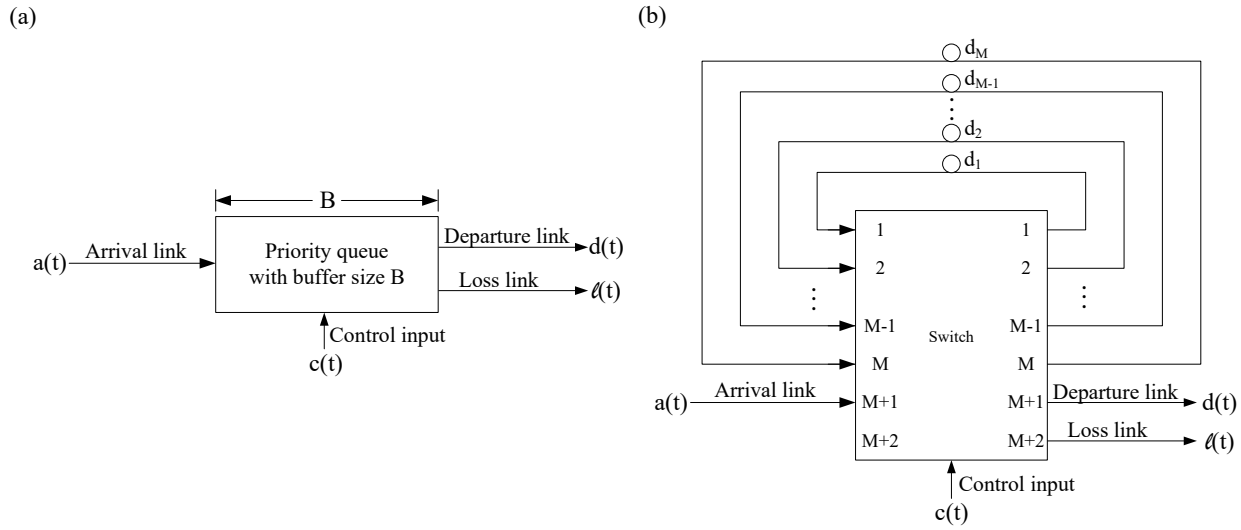

Fig. 1. (a) A priority queue with buffer size B. (b) A construction of an optical priority queue in [13].

packet in the queue has a distinct priority and the relative priority order between any two packets remains unchanged as long as they are in the queue. When there is a departure request from the controller and there are packets in the queue, the packet with the *highest* priority is sent out from the queue through the departure link. When there is a buffer overflow, the packet with the *lowest* priority is dumped from the queue through the loss link. Since packet arrival times and packet departure requests can be arbitrary, and packet priority assignments can also be arbitrary as long as the above-mentioned constraints on packet priorities are satisfied, it is clear that priority queues are very general and include FIFO queues and LIFO queues as special cases. However, this also means that the design of optical priority queues is expected to be more challenging than the other types of optical queues.

The first construction of optical priority queues appeared in [13], in which an optical priority queue with buffer size $O(M^2)$ was constructed by using a feedback system consisting of an optical $(M+2)\times(M+2)$ (bufferless) crossbar switch and M fiber delay lines (see Figure 1(b)). A theoretical upper bound 2^M on the buffer size that can be achieved by using such a feedback system was also given in [13]. The proof in [13] is quite elaborate and a simpler proof was given in [14]. The buffer size $O(M^2)$ achieved in [13] was improved to $O(M^3)$ in [15], and was improved to $O(M^c)$ for any positive integer c in [18]. The constructions in [13]–[15] use a *sorting*-based routing policy, where the packets at the input links of the crossbar switch are first sorted according to their priorities and then routed to the departure link, the loss link, or the fiber delay lines. Such a sorting-based approach only uses the relative priority order among packets at the input links of the crossbar switch, instead of directly using their priorities, to design the routing policy performed by the crossbar switch, and this is the main reason why the buffer sizes achieved in these constructions are limited to polynomial in M.

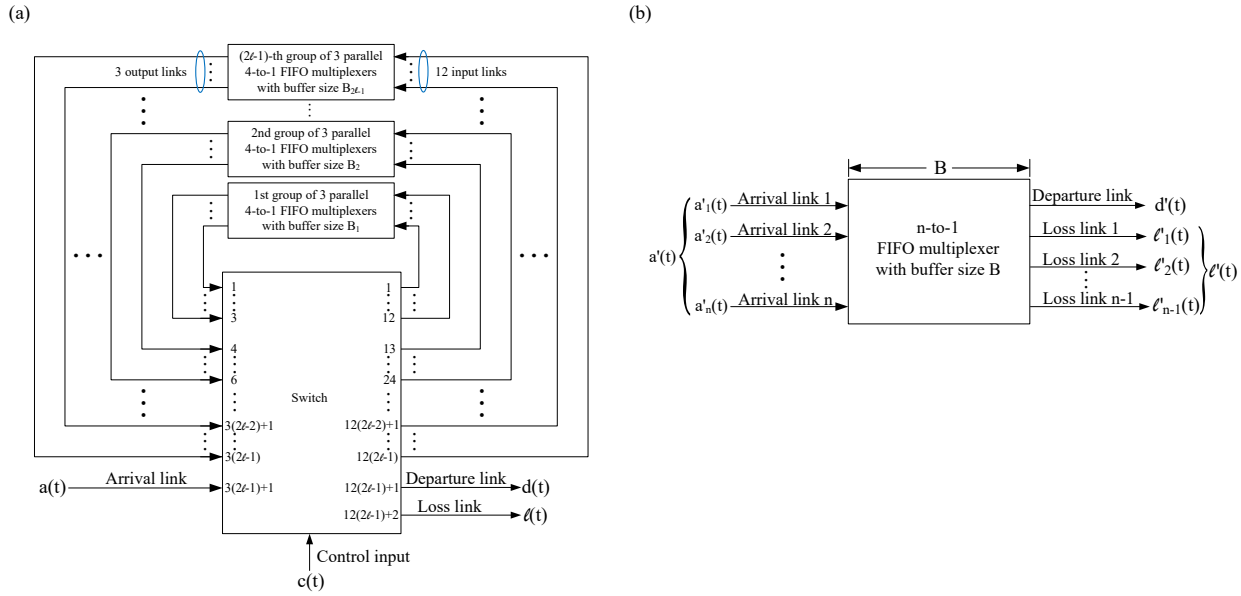

Fig. 2. (a) The feedback system in [19]. (b) An n-to-1 FIFO multiplexer with buffer size B.

To achieve a buffer size beyond polynomial in M, it was proposed in [19] to replace each fiber delay line in Figure 1(b) with a group of three parallel optical 4-to-1 FIFO multiplexers (see Figure 2(a)), and use a simple *priority*-based routing policy that directly uses the priorities of the packets at the input links of the crossbar switch for the routing of packets. An n-to-1 FIFO multiplexer (nFM) with buffer size B has n arrival links, one departure link, and $n-1$ loss links (see Figure 2(b)), where the packet with the earliest arrival time leaves from the departure link whenever there are packets in the nFM, and the packets with the latest arrival times are dumped through the loss links whenever there is a buffer overflow at the nFM. At each slot t, a packet p in the feedback system in Figure 2(a) is associated with a unique positive integer $\tau_p(t)$, called the *tag* of packet p at slot t, to indicate its priority level so that the i^{th}-highest-priority packet in the queue has a tag equal to i. Specifically, if there are $q(t-1)$ packets stored in the buffers of the 4FM's at slot $t-1$ and there are $a(t)$ arrival packets at slot t, then the $q(t-1)+a(t)$ packets in the queue at slot t are assigned tags from 1 to $q(t-1)+a(t)$ in the order of decreasing priority. Furthermore, each group of 4FM's is associated with a unique set of tags, say the i^{th} group of 4FM's is associated with the set Ψ_i of tags for all $i = 1, 2, \ldots, 2\ell-1$. At slot t, if a packet p at the input links of the crossbar switch is not routed to the departure link or the loss link, then it has to be stored in the buffers of the 4FM's, and it is routed to the i^{th} group of 4FM's if $\tau_p(t) \in \Psi_i$ under the priority-based routing policy in [19].

Two problems can arise in [19] as described below. (i) The *no-buffering* problem: We assume that the feedback system in Figure 2(a) is initially empty at slot $t = 0$. At slot $t = 1$, suppose that there is an arrival packet, say packet p, and there is no departure request. Then packet p has to be buffered in the 4FM's at slot $t = 1$. As $\tau_p(1) = 1$ (note that packet p is the only packet

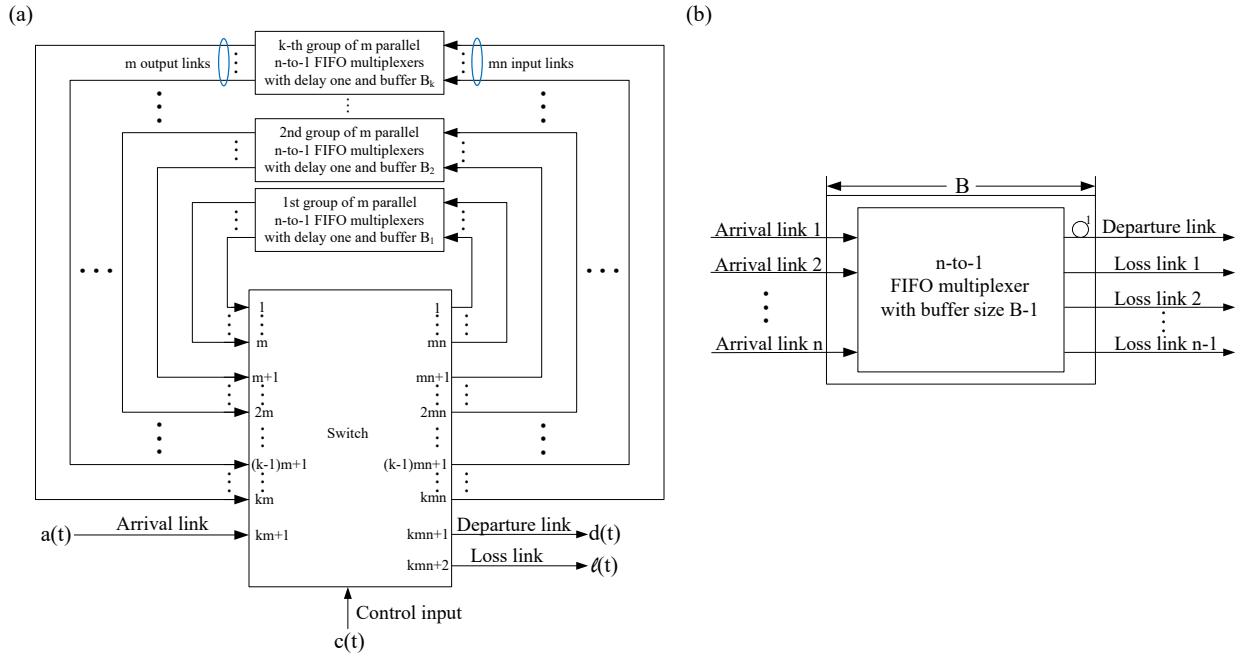

Fig. 3. (a) A construction of an optical priority queue by using an optical $(kmn+2)\times(kmn+2)$ (bufferless) crossbar switch and k groups of m parallel optical n-to-1 FIFO multiplexers with delay one. (b) An n-to-1 FIFO multiplexer with delay one and buffer size B.

in the queue at slot $t=1$) and $\Psi_1=\{1\}$ (according to the assignment of the sets Ψ_i's in [19]), we see that $\tau_p(1)\in\Psi_1$ and hence packet p is routed to the first group of 4FM's at slot $t=1$ under the priority-based routing policy in [19]. Since the 4FM to which packet p is routed is empty when packet p arrives, packet p is immediately sent out from that 4FM and thus it is not successfully buffered at slot $t=1$. This leads to the failure of the constructions in [19]. To fix such a no-buffering problem, in this paper we propose to replace the optical 4FM's in Figure 2(a) with optical 4-to-1 FIFO multiplexers with delay one (see Figure 3(a) with $k=2\ell-1$ for some $\ell\geq 2$, $m=3$, and $n=4$). An optical n-to-1 FIFO multiplexer with delay one (nFM1) and buffer size B is defined as the concatenation of an optical nFM with buffer size $B-1$ and a fiber delay line with delay equal to one, where the departure link of the nFM is connected to the input link of the fiber delay line (see Figure 3(b)). As it takes one slot for a packet to traverse through a fiber delay line with delay equal to one, a packet admitted into an nFM1 is buffered there for at least one slot. This solves the no-buffering problem.

(ii) *Inefficient utilization of buffering capacity*: Assume that we have replaced the optical 4FM's in Figure 2(a) with optical 4FM1's so as to fix the no-buffering problem as mentioned above in (i), and assume that the feedback system is initially empty at slot $t=0$. Consider the case that $\ell=2$ so that $B_1=B_2=B_3=1$, $\Psi_1=\{1\}$, $\Psi_2=\{2,3\}$, and $\Psi_3=\{4\}$ according to the assignment of B_i's and Ψ_i's in [19]. At slot $t=1$, suppose that there is an arrival packet, say packet p_1, and there is no departure request. As already discussed in (i), packet p_1 is routed to

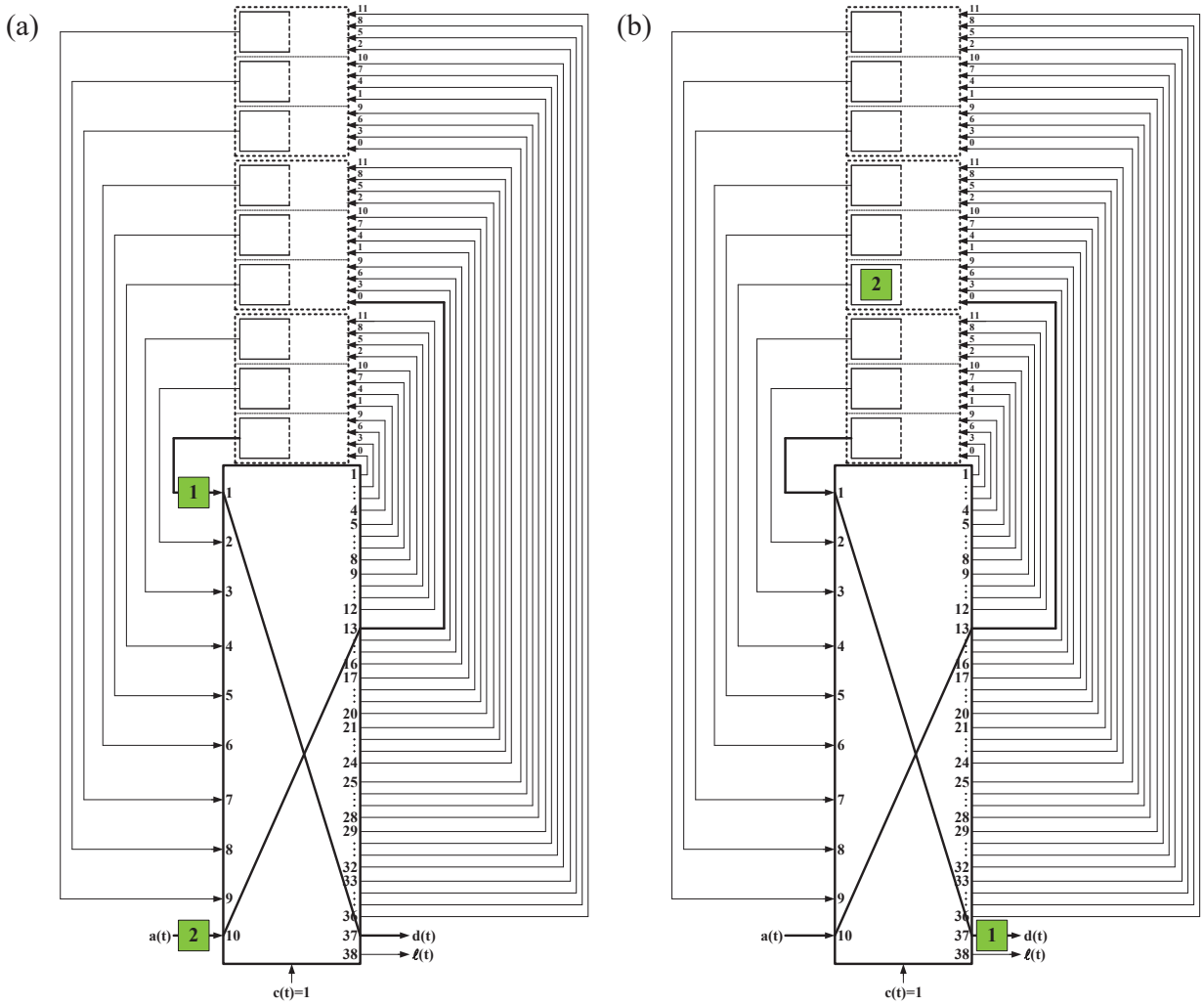

Fig. 4. An illustration of the inefficient utilization of buffering capacity in [19]. (a) At slot $t = 2$, packet p_1 with $\tau_{p_1}(2) = 1$ leaves from the first group of 4FM1's and appears at input link 1 of the crossbar switch, and packet p_2 with $\tau_{p_2}(2) = 2 \in \Psi_2$ arrives from the arrival link. (b) At slot $t = 2$, packet p_1 is routed to the departure link and packet p_2 is routed to the second group of 4FM1's under the priority-based routing policy in [19]. We note that the input links of each group of 4FM1's are numbered from 0 to 11 and this will be used in the description of the round-robin routing policy (R3) in Section II-C.

the first 4FM1 in the first group of 4FM1's at slot $t = 1$. As $B_1 = 1$, packet p_1 will be buffered in that 4FM for one slot and then leave from the first group of 4FM1's and appear at input link 1 of the crossbar switch at slot $t = 2$ (see Figure 4(a)). At slot $t = 2$, suppose that there is another arrival packet, say packet p_2, with priority lower than packet p_1 (see Figure 4(a)), and there is a departure request. Then it is clear that $\tau_{p_1}(2) = 1$ and $\tau_{p_2}(2) = 2$, and hence at slot $t = 2$ packet p_1 is routed to the departure link (as it is the highest-priority packet in the queue) and packet p_2 is routed to the second group of 4FM1's (as $\tau_{p_2}(2) \in \Psi_2$) as shown in Figure 4(b) under the priority-based routing policy in [19]. Since the 4FM1's in the *first* group are empty at slot $t = 2$, their buffers are not used at slot $t = 2$, and this is a waste in the utilization of

buffering capacity. In general, when there is a departure request at slot t, the packet with tag one (i.e., the packet with the highest priority in the queue) is routed to the departure link, and the other packets at the input links of the crossbar switch have tags greater than one and hence are routed to the groups of 4FM1's other than the first group (since $\Psi_1 = \{1\}$). Therefore, the buffers in the first group are not used at slot t.

A simple way to improve this situation is to focus on the packets that have to be stored in the buffers of the 4FM1's. At each slot t, a packet p that has to be stored in the buffers of the 4FM1's is associated with a unique positive integer $\widetilde{\tau}_p(t)$, called the *buffering tag* of packet p at slot t, so that the i^{th}-highest-priority packet among all of the packets that have to be buffered in the 4FM1's has a buffering tag equal to i. Specifically, if there are $q(t-1)$ packets stored in the buffers of the 4FM1's at slot $t-1$ and there are $a(t)$ arrival packets, $d(t)$ departure packets, and $\ell(t)$ loss packets at slot t, then the $q(t-1)+a(t)-d(t)-\ell(t)$ packets that have to be buffered in the 4FM1's at slot t are assigned buffering tags from 1 to $q(t-1)+a(t)-d(t)-\ell(t)$ in the order of decreasing priority. Note that a packet with a smaller buffering tag has a higher priority than a packet with a larger buffering tag. The i^{th} group of 4FM1's is now associated with a set Ψ_i of buffering tags for all $i = 1, 2, \ldots, 2\ell-1$. At slot t, a packet p with buffering tag $\widetilde{\tau}_p(t) \in \Psi_i$ is routed to the i^{th} group of 4FM1's. Since we number the buffering tags starting from 1, the packet with buffering tag equal to 1 is always routed to the first group of 4FM1's (as $1 \in \Psi_1 = \{1\}$) so that the buffers in the first group of 4FM1's are utilized. This improves the utilization of buffering capacity. In the above example, we have $\widetilde{\tau}_{p_2}(2) = 1$ (as packet p_2 is the only packet that has to be buffered in the 4FM1's at slot $t = 2$), and hence packet p_2 is routed to the first group of 4FM1's at slot $t = 2$ (as $\widetilde{\tau}_{p_2}(2) = 1 \in \Psi_1 = \{1\}$).

In this paper, we not only fix the no-buffering problem (by replacing the 4FM's in Figure 2(a) with 4FM1's) and improve the utilization of buffering capacity over that in [19] (by using buffering tags, instead of tags, in the priority-based routing policy), but also extend and generalize the construction in [19] and obtain a much larger class of constructions of optical priority queues. Specifically, we use the feedback system in Figure 3(a) consisting of an optical $(kmn+2) \times (kmn+2)$ (bufferless) crossbar switch and k groups of optical nFM1's, where the i^{th} group has m parallel optical nFM1's with the same buffer size B_i ($B_i \geq 1$) for $i = 1, 2, \ldots, k$. We show in Theorem 7 (see Section III) that the feedback system in Figure 3(a) can be operated as an optical priority queue with buffer size $U_k = \sum_{i=1}^{k} |\Psi_i|$ under the priority-based routing policy (R1)–(R3) (see Section II-C) if $1 \leq s \leq k-1$, where s is a parameter in the conditions (A1)–(A3) (see Section III), $m \geq 1$, and $n, B_1, B_2, \ldots, B_k, |\Psi_1|, |\Psi_2|, \ldots, |\Psi_k|$ satisfy the conditions (A1)–(A3). We note that the construction in [19] is a special case of the constructions in this paper with $s = 1$, $k = 2\ell - 1$ for some $\ell \geq 2$, $m = 3$, $n = 4$, $B_1 = B_{2\ell-1} = 1$, $B_i = B_{2\ell-i} = 2^{i-2}$ for $2 \leq i \leq \ell$, and $|\Psi_i| = |\Psi_{2\ell-i}| = 2^{i-1}$ for $1 \leq i \leq \ell$.

Our constructions are made possible by showing that the highest-priority (resp., lowest-priority) packet is always available at the input links of the crossbar switch whenever there is a departure request and there are packets in the queue (resp., whenever there is a buffer overflow) so that it

can be routed to the departure (resp., loss) link whenever necessary, and by showing that there is no collision and there is no buffer overflow at any nFM1 at any slot so that there is no internal packet loss in the queue at any slot.

The rest of this paper is organized as follows. In Section II, we give more details about priority queues and nFM1's, describe the priority-based routing policy performed by the crossbar switch in Figure 3(a), and derive some basic properties on the buffering tags under our priority-based routing policy. Then we show in Section III that the feedback system in Figure 3(a) can be operated as an optical priority queue under our priority-based routing policy. In Section IV, we perform a complexity analysis for our constructions with maximum buffer sizes and show that a buffer size of $2^{O(\sqrt{\alpha M})}$ can be achieved by using an optical $(M+2) \times (M+2)$ (bufferless) crossbar switch and M fiber delay lines, where α is a constant that depends on the parameters used in our constructions. In Section V, we describe the router buffer sizing problem, present our numerical results, and discuss some feasibility issues. Finally, we conclude this paper in Section VI.

II. Priorities Queues, nFM1's, Priority-Based Routing Policy, and Basic Properties on Buffering Tags

In this paper, we assume that every network element is initially empty at slot $t = 0$. Recall that for the sake of conciseness, we have abbreviated n-to-1 FIFO multiplexer (resp., n-to-1 FIFO multiplexer with delay one) as nFM (resp., nFM1). We have also denoted $\tau_p(t)$ (resp., $\widetilde{\tau}_p(t)$) as the tag (resp. buffering tag) of a packet p in a priority queue at slot t.

Since a packet can be transmitted through a link within a slot, there can be at most one packet in a link at any slot, and hence we can characterize a link by its link state, say a link is in state 1 (resp., 0) at slot t if there is a packet (resp., there is no packet) in the link at slot t.

A. Priorities Queues

For a priority queue with buffer size B as shown in Figure 1(a), we denote $a(t)$, $d(t)$, and $\ell(t)$ as the link states of the arrival link, the departure link, and the loss link, respectively, at slot t. We denote $c(t) = 1$ (resp., $c(t) = 0$) if there is a departure request (resp., there is no departure request) from the controller at slot t. Let $q(t)$ be the number of packets stored in the buffer of the priority queue at slot t.

Then a priority queue with buffer size B is characterized by the following five properties: (P1) *Flow conservation:* Packets arriving from the arrival link are either stored in the buffer or transmitted through the departure link or the loss link. Thus, we have $q(t) = q(t-1) + a(t) - d(t) - \ell(t)$. (P2) *Nonidling:* There is a departure packet at slot t only when there is a departure request from the controller and there are packets in the queue at slot t. Thus, we have $d(t) = 1$ if $c(t) = 1$ and $q(t-1) + a(t) > 0$, and $d(t) = 0$ otherwise. (P3) *Maximum buffer usage:* There is a loss packet at slot t only when there is a buffer overflow at slot t. Thus, we have $\ell(t) = 1$ if $c(t) = 0$, $q(t-1) = B$, and $a(t) = 1$, and $\ell(t) = 0$ otherwise. (P4) *Priority departure:* If

there is a departure packet, say packet p, at slot t, then packet p is the packet with the highest priority in the queue at slot t, i.e., $\tau_p(t) = 1$. (P5) *Priority loss:* If there is a loss packet, say packet p, at slot t, then packet p is the packet with the lowest priority in the queue at slot t, i.e., $\tau_p(t) = B + 1$.

We note that the tag and the buffering tag of a packet in a priority queue can change as time evolves due to the arrivals and departures of packets with priorities higher than that packet (packets with priorities lower than that packet have no effect on the change of its tag or buffering tag). Specifically, consider the scenario that the properties (P4) and (P5) are satisfied at slot t and a packet p in the queue at slot t is not the departure packet (if any) or the loss packet (if any) at slot t so that it has to be buffered in the queue at slot t. Then it is clear that the departure (resp., loss) packet (if any) at slot t has priority higher (resp., lower) than packet p by (P4) (resp., by (P5)), and hence we have

$$\widetilde{\tau}_p(t) = \tau_p(t) - d(t). \tag{1}$$

Now consider the scenario that the properties (P1), (P4), and (P5) are satisfied at slot $t-1$ and a packet p is buffered in the queue at slot $t-1$. Then it is clear that there is no internal packet loss in the queue at slot $t-1$ (by (P1)) and hence we have

$$\tau_p(t) = \tau_p(t-1) - d(t-1) + a_p(t), \tag{2}$$

where $a_p(t)$ is the number of arrival packets at slot t with priorities higher than packet p.

Furthermore, consider the scenario that the property (P1) is satisfied at slot $t-1$ and the properties (P4) and (P5) are satisfied at slots $t-1$ and t, and a packet p is buffered in the queue at slot $t-1$ and has to be buffered in the queue at slot t. Then we see from (1) and (2) (note that we have used (1) twice) that

$$\widetilde{\tau}_p(t) = \widetilde{\tau}_p(t-1) - d(t) + a_p(t), \tag{3}$$

where $a_p(t)$ is the number of arrival packets at slot t with priorities higher than packet p.

B. n-to-1 FIFO Multiplexers and n-to-1 FIFO Multiplexers with Delay One

An nFM with buffer size B is shown in Figure 2(b). To break the tie among packets arriving at the same time, we assume that the arrival links are prioritized so that the priorities of the arrival links are decreasing in the link indices, i.e., packets from arrival links with smaller link indices are regarded as arriving earlier than those from arrival links with larger link indices. We denote $a'_i(t)$ as the link state of arrival link i at slot t for $i = 1, 2, \ldots, n$, denote $d'(t)$ as the link state of the departure link at slot t, and denote $\ell'_i(t)$ as the link state of loss link i at slot t for $i = 1, 2, \ldots, n-1$. Let $a'(t) = \sum_{i=1}^{n} a'_i(t)$ and $\ell'(t) = \sum_{i=1}^{n-1} \ell'_i(t)$ be the number of packets arriving from the arrival links and the number of packets dumped through the loss links, respectively, at slot t. Let $q'(t)$ be the number of packets buffered in the nFM at slot t.

Then an nFM with buffer size B is characterized by the following five properties: (M1) *Flow conservation:* This property is the same as property (P1). Thus, we have $q'(t) = q'(t-1) + a'(t) -$

$d'(t) - \ell'(t)$. (M2) *Nonidling:* There is a departure packet at slot t whenever there are packets in the queue at slot t. Thus, we have $d'(t) = 1$ if $q'(t-1) + a'(t) > 0$, and $d'(t) = 0$ otherwise. (M3) *Maximum buffer usage:* There is a loss packet at slot t only when there is a buffer overflow at slot t. Thus, we have $\ell'(t) = q'(t-1) + a'(t) - 1 - B$ if $q'(t-1) + a'(t) - 1 > B$, and $\ell'(t) = 0$ otherwise. (M4) *FIFO departure:* Packets depart in the FIFO order. (M5) *FIFO loss with prioritized loss links:* If there are loss packets at slot t, i.e., $\ell'(t) > 0$, then the loss packets are the *latest* $\ell'(t)$ arrival packets at slot t and they are dumped through loss links $1, 2, \ldots, \ell'(t)$ in the order of increasing arrival link indices.

An nFM1 with buffer size B is defined as the concatenation of an optical nFM with buffer size $B-1$ and a fiber delay line with delay equal to one as shown in Figure 3(b). We make the following remark on nFM1's that will be useful later in this paper.

Remark 1 *(i) From the properties (M2) and (M4), we can see that a packet admitted into an nFM with buffer size $B-1$ is buffered there for at most $B-1$ slots. Therefore, it is clear from Figure 3(b) that a packet admitted into an nFM1 with buffer size B is buffered there for at least one slot and at most B slots. It follows that a packet admitted into an nFM1 with buffer size $B = 1$ is buffered there for exactly one slot.*

(ii) When there are packets buffered in an nFM1 with buffer size B as shown in Figure 3(b), it is easy to see from the property (M2) that one of those packets must be buffered in the fiber delay line with delay one. Thus, the buffering capacity of the fiber with delay one is fully utilized. As an nFM with buffer size $B-1$ has the capability of buffering $B-1$ packets, it then follows that the effective buffering capacity of such a concatenation in Figure 3(b) is B.

C. The Priority-Based Routing Policy

As mentioned in Section I, each group of nFM1's in Figure 3(a) is associated with a unique set of buffering tags. Specifically, the i^{th} group of nFM1's is associated with the set Ψ_i of buffering tags for $i = 1, 2, \ldots, k$ as described below. Let U_k be the targeted buffer size of the optical priority queue in our construction. Partition the set $\Psi = \{1, 2, \ldots, U_k\}$ of buffering tags into k pairwise disjoint nonempty subsets $\Psi_i = \{U_{i-1} + 1, U_{i-1} + 2, \ldots, U_i\}$, $i = 1, 2, \ldots, k$, where

$$U_0 = 0 < U_1 < U_2 < \cdots < U_k. \tag{4}$$

It is clear that $|\Psi_i| = U_i - U_{i-1}$ for $i = 1, 2, \ldots, k$ and $U_i = \sum_{j=1}^{i} |\Psi_j|$ for $i = 1, 2, \ldots, k$. Let $L_i = U_{i-1} + 1$ so that we can write Ψ_i as $\Psi_i = \{L_i, L_i + 1, \ldots, U_i\}$ for $i = 1, 2, \ldots, k$. Note that $L_1 = U_0 + 1 = 1$ and we have from (4) that $L_i \leq U_i$ for $i = 1, 2, \ldots, k$.

Then the crossbar switch in Figure 3(a) is operated according to the following priority-based routing policy at all slots $t \geq 1$.

(R1) Departure packets: If there is a departure request from the controller and there are packets in the queue at slot t, i.e., $c(t) = 1$ and $q(t-1) + a(t) > 0$, then the highest-priority packet (if any) among all of the packets from the arrival link or the m output links of the *first*

group of nFM1's is routed to the departure link at slot t. Otherwise, no packet is routed to the departure link at slot t.

(R2) Loss packets: If there is a buffer overflow at slot t, i.e., $c(t) = 0$, $q(t-1) = U_k$, and $a(t) = 1$, then the lowest-priority packet (if any) among all of the packets from the arrival link or the m output links of the *last* group of nFM1's is routed to the loss link at slot t. Otherwise, no packet is routed to the loss link at slot t.

(R3) Round-robin routing at the k groups of nFM1's: A packet p at the input links of the crossbar switch that has to be buffered in the queue (i.e., it is not routed to the departure link according to (R1) or the loss link according to (R2)) and has $\widetilde{\tau}_p(t) \in \Psi_i$ is routed to the i^{th} group of nFM1's. Furthermore, packets routed to a group of nFM1's are distributed to the nFM1's in that group in a *round-robin* fashion so that load balancing among the nFM1's in that group can be achieved and hence the buffering capacity of the nFM1's can be fully utilized. Specifically, the round-robin routing is described as follows. Consider the i^{th} group, where $1 \le i \le k$. For ease of presentation, we call arrival link ℓ of the j^{th} nFM1 in the i^{th} group the $((\ell-1)m+j-1)^{\text{th}}$ input link of the i^{th} group for $j = 1, 2, \ldots, m$ and $\ell = 1, 2, \ldots, n$. As such, the inputs of the i^{th} group are numbered from 0 to $mn-1$ (see Figure 4 for an illustration). Let $u_i(0) = 0$. At slot t, if there are $r_i(t)$ packets routed to the i^{th} group, then they are routed to the $(u_i(t-1) \mod mn)^{\text{th}}$, $((u_i(t-1)+1) \mod mn)^{\text{th}}$, …, $((u_i(t-1)+r_i(t)-1) \mod mn)^{\text{th}}$ input links of the i^{th} group in the order of increasing buffering tags, and we update $u_i(t)$ as $u_i(t) = (u_i(t-1) + r_i(t)) \mod mn$. It is clear that $u_i(t)$ is the index of the input link of the i^{th} group that will be firstly used by the packets routed to the i^{th} group at slot t.

Remark 2 *(i) We will show in the proof of Theorem 7 that the following four conditions are satisfied at all slots $t \ge 1$ under the priority-based routing policy (R1)–(R3): (C1)* Highest-priority packet availability *condition: If there is a departure request and there are packets in the queue at slot t, then the packet with the highest priority in the queue at slot t is from the arrival link or the m output links of the first group of nFM1's. (C2)* Lowest-priority packet availability *condition: If there is a buffer overflow at slot t, then the packet with the lowest priority in the queue at slot t is from the arrival link or the m output links of the last group of nFM1's. (C3)* Collision-free *condition: There is at most one packet routed to any input link of any nFM1 at slot t. (C4)* No buffer overflow *condition: There is no buffer overflow at any nFM1 at slot t.*

(ii) If the condition (C1) (resp., (C2)) is satisfied at slot t, then we see from the routing policy (R1) (resp., (R2)) that the properties (P2) and (P4) (resp., (P3) and (P5)) are satisfied at slot t. If the conditions (C3) and (C4) are satisfied at slot t, then there is no packet loss at any nFM1 so that there is no internal packet loss in the feedback system in Figure 3(a) at slot t, and hence the property (P1) is satisfied at slot t. Therefore, if the conditions (C1)–(C4) are satisfied at slot t, then the feedback system in Figure 3(a) can be operated as a priority queue with buffer size U_k at slot t.

D. Basic Properties on Buffering Tags

In this subsection, we derive some basic properties on buffering tags that will be used in the proof of our constructions of optical priority queues in Section III.

We first derive two basic properties on the change of buffering tags in a slot under our priority-based routing policy. The first property says that the buffering tag of a packet can only increase (resp., decrease) by at most one in a slot under our priority-based routing policy, which is a direct result of (3) and the fact that there is at most one arrival (resp., departure) packet with priority higher than that packet in a slot.

Theorem 3 *Assume that the feedback system in Figure 3(a) is operated under the routing policy (R1)–(R3) at all slots, the property (P1) is satisfied up to slot $t-1$, and the priorities (P4) and (P5) are satisfied up to slot t. Suppose that a packet p is buffered in the feedback system at slot $t-1$ and has to be buffered in the feedback system at slot t. Then we have*

$$-1 \leq \widetilde{\tau}_p(t) - \widetilde{\tau}_p(t-1) \leq 1. \tag{5}$$

The second property says that the difference between the buffering tag of a lower-priority packet and the buffering tag of a higher-priority packet cannot decrease and can only increase by at most one in a slot under our priority-based routing policy, which is a direct consequence of (3) and the fact that there is at most one arrival packet with priority lower than the higher-priority packet but higher than the lower-priority packet in a slot.

Theorem 4 *Assume that the feedback system in Figure 3(a) is operated under the routing policy (R1)–(R3) at all slots, the property (P1) is satisfied up to slot $t-1$, and the priorities (P4) and (P5) are satisfied up to slot t. Suppose that two packets, say packet p_1 and packet p_2, are buffered in the feedback system at slot $t-1$ and have to be buffered in the feedback system at slot t, where packet p_1 has higher priority than packet p_2, i.e., $\widetilde{\tau}_{p_1}(t-1) < \widetilde{\tau}_{p_2}(t-1)$. Then we have*

$$0 \leq [\widetilde{\tau}_{p_2}(t) - \widetilde{\tau}_{p_1}(t)] - [\widetilde{\tau}_{p_2}(t-1) - \widetilde{\tau}_{p_1}(t-1)] \leq 1. \tag{6}$$

In the following, we derive two basic properties on the buffering tags of packets buffered in or routed to each group of nFM1's under our priority-based routing policy. We first use Theorem 3 to derive the range of the buffering tags of packets buffered in each group of nFM1's under our priority-based routing policy.

Theorem 5 *Assume that the feedback system in Figure 3(a) is operated under the routing policy (R1)–(R3) at all slots, the property (P1) is satisfied up to slot $t-1$, and the priorities (P4) and (P5) are satisfied up to slot t. Suppose that a packet p is buffered in the i^{th} group of nFM1's at slot t for some $1 \leq i \leq k$. Then we have*

$$L_i - B_i + 1 \leq \widetilde{\tau}_p(t) \leq U_i + B_i - 1. \tag{7}$$

Proof. Let t' be the slot that packet p is routed to the i^{th} group of nFM1's for the last time before or at slot t, say packet p is routed to the j^{th} nFM1 in the i^{th} group at slot t' for some

$1 \le j \le m$. Since packet p is buffered in the i^{th} group of nFM1's at slot t, it is clear from the definition of t' that packet p is admitted into the j^{th} nFM1 in the i^{th} group at slot t' and buffered there at slots $t', t'+1, \ldots, t$. As we also know from Remark 1(i) that after packet p is admitted into the the j^{th} nFM1 in the i^{th} group at slot t', it can be buffered there for at most B_i slots, we easily deduce that

$$t \le t' + B_i - 1. \tag{8}$$

Now write $\widetilde{\tau}_p(t)$ as

$$\widetilde{\tau}_p(t) = \widetilde{\tau}_p(t') + \sum_{\ell=1}^{t-t'} (\widetilde{\tau}_p(t'+\ell) - \widetilde{\tau}_p(t'+\ell-1)). \tag{9}$$

It then follows from (9), $\widetilde{\tau}_p(t') \in \Psi_i = \{L_i, L_i+1, \ldots, U_i\}$ (according to the routing policy (R3)), Theorem 3, and $t - t' \le B_i - 1$ in (8) that $\widetilde{\tau}_p(t) \le U_i + (t-t') \cdot 1 \le U_i + B_i - 1$ and $\widetilde{\tau}_p(t) \ge L_i - (t-t') \cdot 1 \ge L_i - B_i + 1$. ■

Now we use Theorem 3 and Theorem 4 to derive an upper bound on the difference between the buffering tags of two packets that are buffered in or routed to each group of nFM1's, which in turn gives an upper bound on the number of packets buffered in or routed to each group of nFM1's under our priority-based routing policy.

Theorem 6 *Assume that the feedback system in Figure 3(a) is operated under the routing policy (R1)–(R3) at all slots, the property (P1) is satisfied up to slot $t-1$, and the priorities (P4) and (P5) are satisfied up to slot t. Suppose that two packets, say packet p_1 and packet p_2, are buffered in or routed to the i^{th} group of nFM1's at slot t for some $1 \le i \le k$. Then we have*

$$|\widetilde{\tau}_{p_1}(t) - \widetilde{\tau}_{p_2}(t)| \le |\Psi_i| + B_i - 2. \tag{10}$$

Therefore, there are at most $|\Psi_i| + B_i - 1$ packets buffered in or routed to the i^{th} group of nFM1's at slot t.

Proof. Let t_1 (resp., t_2) be the slot that packet p_1 (resp., packet p_2) is routed to the i^{th} group of nFM1's for the last time before or at slot t. Then we have $\widetilde{\tau}_{p_1}(t_1), \widetilde{\tau}_{p_2}(t_2) \in \Psi_i = \{L_i, L_i + 1, \ldots, U_i\}$ (according to the routing policy (R3)), and it follows that

$$|\widetilde{\tau}_{p_1}(t_1) - \widetilde{\tau}_{p_2}(t_2)| \le U_i - L_i = |\Psi_i| - 1. \tag{11}$$

Assume without loss of generality that $t_1 \le t_2$. Since $t_1 \le t$, we consider the two cases $t_1 = t$ and $t_1 < t$ separately.

Case 1: $t_1 = t$. In this case, we see from $t = t_1 \le t_2 \le t$ that $t_2 = t$. Thus, we have from (11) and $B_i \ge 1$ that

$$|\widetilde{\tau}_{p_1}(t) - \widetilde{\tau}_{p_2}(t)| = |\widetilde{\tau}_{p_1}(t_1) - \widetilde{\tau}_{p_2}(t_2)| \le |\Psi_i| - 1 \le |\Psi_i| + B_i - 2.$$

Case 2: $t_1 < t$. In this case, packet p_1 is not routed to the i^{th} group of nFM1's at slot t (according to the definition of t_1) and hence it must be buffered in the i^{th} group of nFM1's at slot t. It then follows from the argument leading to (8) that $t \le t_1 + B_i - 1$. Therefore, we have

$$\begin{aligned}
&|\widetilde{\tau}_{p_1}(t) - \widetilde{\tau}_{p_2}(t)| \\
&= \left| \widetilde{\tau}_{p_1}(t_2) - \widetilde{\tau}_{p_2}(t_2) + \sum_{\ell=1}^{t-t_2} \left[(\widetilde{\tau}_{p_1}(t_2+\ell) - \widetilde{\tau}_{p_2}(t_2+\ell)) - (\widetilde{\tau}_{p_1}(t_2+\ell-1) - \widetilde{\tau}_{p_2}(t_2+\ell-1)) \right] \right| \\
&\le |\widetilde{\tau}_{p_1}(t_2) - \widetilde{\tau}_{p_2}(t_2)| + (t-t_2) \cdot 1 \\
&= \left| \widetilde{\tau}_{p_1}(t_1) - \widetilde{\tau}_{p_2}(t_2) + \sum_{\ell=1}^{t_2-t_1} (\widetilde{\tau}_{p_1}(t_1+\ell) - \widetilde{\tau}_{p_1}(t_1+\ell-1)) \right| + t - t_2 \\
&\le |\Psi_i| - 1 + (t_2 - t_1) \cdot 1 + t - t_2 \\
&\le |\Psi_i| + B_i - 2,
\end{aligned}$$

where the first inequality follows from Theorem 4, the second inequality follows from (11) and Theorem 3, and the third inequality follows from $t \le t_1 + B_i - 1$. ■

III. Constructions of Optical Priority Queues

In this section, we will use the basic properties on buffering tags obtained in Section II-D to show that the feedback system in Figure 3(a) can be operated as an optical priority queue with buffer size U_k under the routing policy (R1)–(R3) if $1 \le s \le k-1$, $m \ge 1$, and $n, B_1, B_2, \ldots, B_k, |\Psi_1|, |\Psi_2|, \ldots, |\Psi_k|$ satisfy the following conditions (A1)–(A3):

(A1) $n \ge \min\{2s+1, k\} + 1$.

(A2) $B_1 = B_k = 1$, $B_i \ge 1$ for $i = 2, 3, \ldots, k-1$,

$$B_i \le \begin{cases} U_{i-1}, \text{ if } 2 \le i \le s+1, \\ U_{i-1} - U_{i-s-1}, \text{ if } s+2 \le i \le k, \end{cases}$$

and

$$B_i \le \begin{cases} U_{i+s} - U_i, \text{ if } 1 \le i \le k-s-1, \\ U_k - U_i, \text{ if } k-s \le i \le k-1. \end{cases}$$

(Recall from Section II-C that $U_i = \sum_{j=1}^{i} |\Psi_j|$ for $i = 1, 2, \ldots, k$.)

(A3) $1 \le |\Psi_i| \le (m-1)B_i + 1$ for $i = 1, 2, \ldots, k$.

Theorem 7 *Assume that the feedback system in Figure 3(a) is operated under the routing policy (R1)–(R3) at all slots. Suppose that* $1 \le s \le k-1$, $m \ge 1$, *and* $n, B_1, B_2, \ldots, B_k, |\Psi_1|, |\Psi_2|, \ldots, |\Psi_k|$ *satisfy the conditions (A1)–(A3). Then the feedback system in Figure 3(a) can be operated as an optical priority queue with buffer size* U_k *at all slots* $t \ge 1$.

Remark 8 *It is easy to check that when* $s = 1$, $k = 2\ell - 1$ *for some* $\ell \ge 2$, *and* $m = 3$, *the choice* $n = 4$, $B_1 = B_{2\ell-1} = 1$, $B_i = B_{2\ell-i} = 2^{i-2}$ *for* $2 \le i \le \ell$, *and* $|\Psi_i| = |\Psi_{2\ell-i}| = 2^{i-1}$ *for*

$1 \leq i \leq \ell$ given in [19] satisfies the conditions (A1)–(A3). Therefore, the construction in [19] indeed is a special case of our constructions in Theorem 7 as mentioned in Section I.

Before we present the proof of Theorem 7, we give the intuitive idea behind our constructions. Since the design of the delays of the fiber delay lines in the SDL constructions of the nFM1's adopted in this paper (see Section IV) is determined by the buffer sizes $B_1, B_2, \ldots, B_k$ of the nFM1's, and the design of the routing policy performed by the optical crossbar switches is determined by the sets $\Psi_1, \Psi_2, \ldots, \Psi_k$ of buffering tags under our priority-based routing policy, it is clear that the design of the buffer sizes $B_1, B_2, \ldots, B_k$ and the design of the sets $\Psi_1, \Psi_2, \ldots, \Psi_k$ of buffering tags are closely related and highly coupled as mentioned in Section I.

The idea behind the conditions (A1)–(A3) in our constructions can be roughly described as follows (the details are given in the proof of Theorem 7):

(i) The condition (A2) says that $B_i \leq \sum_{j=1}^{i-1} |\Psi_j|$ for $2 \leq i \leq s+1$, $B_i \leq \sum_{j=i-s}^{i-1} |\Psi_j|$ for $s+2 \leq i \leq k$, $B_i \leq \sum_{j=i+1}^{i+s} |\Psi_j|$ for $1 \leq i \leq k-s-1$, and $B_i \leq \sum_{j=i+1}^{k} |\Psi_j|$ for $k-s \leq i \leq k-1$, namely, B_i is no greater than the sum of the $|\Psi_j|$'s of at most s of its neighboring groups. Note that a packet p may be buffered in an nFM1 in the i^{th} group for up to B_i slots (by Remark 1(i)), and its tag can change as time evolves. If B_i is too large, i.e., greater than such a sum, then packet p may still be buffered in the i^{th} group when its tag decreases to 1 (resp., increases to $U_k + 1$) and there is a departure request (resp., there is a buffer overflow). Therefore, packet p cannot be routed to the departure (resp. loss) link so that we cannot successfully construct an optical priority queue in such a case. In the proof of Theorem 7, we show that such a situation cannot happen and the conditions (C1) and (C2) can be satisfied if the condition (A2) is satisfied.

(ii) Since a packet p may be buffered in an nFM1 in the i^{th} group for up to B_i slots, its buffering tag can change by at most B_i when it leaves from the i^{th} group of nFM1's (by Theorem 3). As we know from the condition (A2) that B_i is no greater than the sum of the $|\Psi_j|$'s of at most s of its neighboring groups, the buffering tag of packet p can only belong to Ψ_j for $\max\{i-s, 1\} \leq j \leq \min\{i+s, k\}$ when packet p leaves from the i^{th} group of nFM1's. Thus, when packet p leaves from the i^{th} group of nFM1's, it can only be routed to the i^{th} group itself or at most $2s$ of its neighboring groups. As a result, the packets routed to a group of nFM1's can only come from the arrival link, or the output links of that group itself or at most $2s$ of its neighboring groups, and this limits the number of packets that can be routed to that group. The condition (A1) then guarantees that n is large enough so that there are enough input links in any group of nFM1's to accommodate the packets routed to that group. Thus, the collision-free condition (C3) can be satisfied if the conditions (A1) and (A2) are satisfied.

(iii) Finally, the condition (A3) says that $|\Psi_i|$ is at most $(m-1)B_i + 1$. If $|\Psi_i|$ is greater than $(m-1)B_i + 1$, then we see from Theorem 6 that there can be more than mB_i packets buffered in or routed to the i^{th} group of nFM1's. Therefore, there are more than B_i packets buffered in or routed to some nFM1 in the i^{th} group, so that there is a buffer overflow at that nFM1. In the

proof of Theorem 7, we show that the no buffer overflow condition (C4) can be satisfied if the condition (A3) is satisfied.

Proof. (Proof of Theorem 7) We will prove this theorem by induction on slot t. Recall that we have assumed that the feedback system in Figure 3(a) is initially empty at slot $t = 0$ and hence we have $q(0) = 0$. First consider slot $t = 1$. As $q(0) = 0$, it is clear that there are $a(1)$ packets in the queue at slot $t = 1$ and they are the arrival packets from the arrival link. Thus, the conditions (C1) and (C2) are trivially satisfied at slot $t = 1$. As $a(1) \leq 1$, it is clear that there is at most one packet routed to any nFM1 at slot $t = 1$. Thus, the conditions (C3) and (C4) are also satisfied at slot $t = 1$. Therefore, it follows from Remark 2(ii) that the feedback system in Figure 3(a) can be operated as an optical priority queue with buffer size U_k at slot $t = 1$.

Now assume as the induction hypothesis that the feedback system in Figure 3(a) can be operated as an optical priority queue with buffer size U_k up to slot $t - 1$, i.e., the properties (P1)–(P5) (with $B = U_k$) are satisfied up to slot $t-1$, for some $t-1 \geq 1$. Therefore, if a packet p is buffered in the queue at slot $t - 1$, then we have from (2) and (1) that

$$\tau_p(t) = \tau_p(t-1) - d(t-1) + a_p(t) = \widetilde{\tau}_p(t-1) + a_p(t), \tag{12}$$

where $a_p(t)$ is the number of arrival packets at slot t with priorities higher than packet p.

In the following, we will show that the conditions (C1)–(C4) are satisfied at slot t. It then follows from Remark 2(ii) that the feedback system in Figure 3(a) can be operated as an optical priority queue with buffer size U_k at slot t, and the induction is completed.

(i) *The highest-priority packet availability condition (C1) is satisfied at slot t.* Suppose that there is a departure request from the controller and there are packets in the queue at slot t, i.e., $c(t) = 1$ and $q(t-1) + a(t) > 0$. We will use Theorem 5 and (A2) to show that the packet with the highest priority in the queue at slot t is from the arrival link or the m output links of the first group of nFM1's so that the condition (C1) is satisfied at slot t.

Let packet p be the packet with the highest priority in the queue at slot t, i.e., $\tau_p(t) = 1$. If packet p is an arrival packet at slot t, then we are done. So assume that packet p is not an arrival packet at slot t. Then packet p must be stored in the buffer of the queue at slot $t - 1$. Let $a_p(t)$ be the number of arrival packets at slot t with priorities higher than packet p. Then it is clear from $\widetilde{\tau}_p(t-1) \geq 1$ and $\widetilde{\tau}_p(t-1) \leq \tau_p(t) = 1$ (by using $a_p(t) \geq 0$ in (12)) that

$$\widetilde{\tau}_p(t-1) = 1. \tag{13}$$

From $B_1 = 1$ in (A2), $L_1 = 1$, and $U_1 \geq L_1$, we have

$$L_1 - B_1 + 1 = L_1 = 1 \text{ and } U_1 + B_1 - 1 = U_1 \geq L_1 = 1. \tag{14}$$

From $L_i = U_{i-1} + 1$, (A2), the monotonicity of the U_i's in (4), and $U_0 = 0$, we also have

$$L_i - B_i + 1 = (U_{i-1} + 1) - B_i + 1 \geq \begin{cases} 2, \text{ if } 2 \leq i \leq s+1, \\ U_{i-s-1} + 2 > U_0 + 2 = 2, \text{ if } s+2 \leq i \leq k. \end{cases} \tag{15}$$

Therefore, we see from Theorem 5 (for slot $t-1$) and (13)–(15) that packet p must be buffered in the first group of nFM1's at slot $t-1$. As $B_1 = 1$, it follows from Remark 1(i) that packet p is buffered there for exactly one slot and then leaves from the first group of nFM1's at slot t.

(ii) *The lowest-priority packet availability condition (C2) is satisfied at slot t.* Suppose that there is a buffer overflow at slot t, i.e., $c(t) = 0$, $q(t-1) = U_k$, and $a(t) = 1$. We will use Theorem 5 and (A2) to show that the packet with the lowest priority in the queue at slot t is from the arrival link or the m output links of the last group of nFM1's so that the condition (C2) is satisfied at slot t.

Let packet p be the packet with the lowest priority in the queue at slot t, i.e., $\tau_p(t) = U_k + 1$. If packet p is an arrival packet at slot t, then we are done. So assume that packet p is not an arrival packet at slot t. Then packet p must be stored in the buffer of the queue at slot $t-1$. Let $a_p(t)$ be the number of arrival packets at slot t with priorities higher than packet p. Then it is clear from $\widetilde{\tau}_p(t-1) \le U_k$ (from the induction hypothesis we know that there are at most U_k packets buffered in the queue at slot $t-1$) and $\widetilde{\tau}_p(t-1) \ge \tau_p(t) - 1 = U_k$ (by using $a_p(t) \le 1$ in (12)) that

$$\widetilde{\tau}_p(t-1) = U_k. \tag{16}$$

From $B_k = 1$ in (A2) and $L_k \le U_k$, we have

$$L_k - B_k + 1 = L_k \le U_k \text{ and } U_k + B_k - 1 = U_k. \tag{17}$$

From (A2) and the monotonicity of the U_i's in (4), we also have

$$U_i + B_i - 1 \le \begin{cases} U_{i+s} - 1 < U_k - 1, \text{ if } 1 \le i \le k-s-1, \\ U_k - 1, \text{ if } k-s \le i \le k-1. \end{cases} \tag{18}$$

Therefore, we see from Theorem 5 (for slot $t-1$) and (16)–(18) that packet p must be buffered in the last group, i.e., the k^{th} group, of nFM1's at slot $t-1$. As $B_k = 1$, it follows from Remark 1(i) that packet p is buffered there for exactly one slot and then leaves from the last group of nFM1's at slot t.

(iii) *The collision-free condition (C3) is satisfied at slot t.* Since we have already shown in (i) and (ii) above that the conditions (C1) and (C2) are satisfied at slot t, we see from Remark 2(ii) that the properties (P4) and (P5) are satisfied at slot t. Thus, we now know from the induction hypothesis that the property (P1) is satisfied up to slot $t-1$, and the priorities (P4) and (P5) are satisfied up to slot t. We will then use Theorem 3, Theorem 5, and (A2) to show that there are at most $m \cdot \min\{2s+1, k\} + 1$ packets routed to any group of nFM1's at slot t. Note that there are mn input links at any group of nFM1's and we have from (A1) that $mn \ge m(\min\{2s+1, k\}+1) \ge m \cdot \min\{2s+1, k\}+1$. As such, it follows from the round-robin routing policy (R3) that there is at most one packet routed to any input link of any nFM1 at slot t. Therefore, the condition (C3) is satisfied at slot t.

It remains to show that there are at most $m \cdot \min\{2s+1, k\} + 1$ packets routed to any group of nFM1's at slot t. Consider a packet, say packet p, that is buffered in the i^{th} group of nFM1's

at slot $t-1$, leaves from the i^{th} group of nFM1's at slot t, and has to be stored in the buffer of the queue at slot t, where $1 \le i \le k$. If $s+2 \le i \le k$, then we have

$$\begin{aligned}\widetilde{\tau}_p(t) - L_{i-s} &\ge (\widetilde{\tau}_p(t-1)-1) - L_{i-s} \ge (L_i - B_i) - L_{i-s} \\ &= (U_{i-1} + 1 - B_i) - (U_{i-s-1}+1) \ge 0, \end{aligned} \tag{19}$$

where the first inequality follows from Theorem 3, the second inequality follows from Theorem 5, and the third inequality follows from (A2). Similarly, if $1 \le i \le k-s-1$, then we also have from Theorem 3, Theorem 5, and (A2) that

$$\widetilde{\tau}_p(t) - U_{i+s} \le (\widetilde{\tau}_p(t-1)+1) - U_{i+s} \le (U_i + B_i) - U_{i+s} \le 0. \tag{20}$$

Thus, we see from (19) and (20) that $\widetilde{\tau}_p(t) \in \Psi_j$ for some $\max\{i-s,1\} \le j \le \min\{i+s,k\}$. It then follows from the routing policy (R3) that packet p can only be routed to the j^{th} group for some $\max\{i-s,1\} \le j \le \min\{i+s,k\}$.

As a result, we can see that the packets routed to the i^{th} group of nFM1's at slot t can only come from the arrival link or the output links of the j^{th} group for some $\max\{i-s,1\} \le j \le \min\{i+s,k\}$. In other words, the packets routed to any group of nFM1's can only come from the arrival link or the output links of at most $\min\{2s+1,k\}$ of groups. As each group has m nFM1's, we conclude that there are at most $m \cdot \min\{2s+1,k\}+1$ packets routed to any group of nFM1's at slot t.

(iv) *The no buffer overflow condition (C4) is satisfied at slot t.* We will use Theorem 6 and (A3) to show that there is no buffer overflow at any nFM1 at slot t. Therefore, the condition (C4) is satisfied at slot t.

Consider the i^{th} group of nFM1's, where $1 \le i \le k$. Let $q'_{i,j}(t')$ (resp., $a'_{i,j}(t')$) be the number of packets buffered in (resp., the number of packets routed to) the j^{th} nFM1 in the i^{th} group at slot t' for $j = 1,2,\ldots,m$ and $t' = 1,2,\ldots$. Consider the j^{th} nFM1 in the i^{th} group, where $1 \le j \le m$. As we know from Remark 1(ii) that there is always a packet buffered in the fiber with delay one in Figure 3(b) (with $B = B_i$) whenever there are packets buffered in the j^{th} nFM1, it follows that among the $q'_{i,j}(t-1)$ packets buffered in the j^{th} nFM1 at slot $t-1$, there are $\min\{q'_{i,j}(t-1),1\}$ of them departing from the j^{th} nFM1 at slot t and the rest of the $(q'_{i,j}(t-1)-1)^+$ packets remain buffered there at slot t, where $x^+ = \max\{x,0\}$. Thus, the number of packets buffered in or routed to the j^{th} nFM1 at slot t is given by $(q'_{i,j}(t-1)-1)^+ + a'_{i,j}(t)$. Therefore, we have from Theorem 6 and (A3) that

$$\sum_{j=1}^{m} ((q'_{i,j}(t-1)-1)^+ + a'_{i,j}(t)) \le |\Psi_i| + B_i - 1 \le mB_i. \tag{21}$$

Since the m nFM1's in the i^{th} group are evenly loaded (according to the round-robin routing policy (R3)) and evenly served (according to Remark 1(ii)), and there is no buffer overflow at any nFM1 up to slot $t-1$ (according to the induction hypothesis), it is a direct result of the *join-the-shortest-queue* and *serve-the-longest-queue policy* in queueing theory that the virtual queue

lengths $(q'_{i,j}(t-1)-1)^+ + a'_{i,j}(t)$, $j = 1, 2, \ldots, m$, of the m nFM1's in the i^{th} group differ by at most one (this fact can also be easily proved by induction on slot t as in the proof of Lemma 11 in [19]). As such, we deduce from (21) that $(q'_{i,j}(t-1)-1)^+ + a'_{i,j}(t) \leq \lceil (mB_i)/m \rceil = B_i$ for $j = 1, 2, \ldots, m$. In other words, there are at most B_i packets buffered in or routed to the j^{th} nFM1 at slot t for $j = 1, 2, \ldots, m$. Therefore, there is no buffer overflow at the j^{th} nFM1 in the i^{th} group at slot t for all $j = 1, 2, \ldots, m$. ■

IV. Complexity Analysis For Constructions With Maximum Buffer Sizes

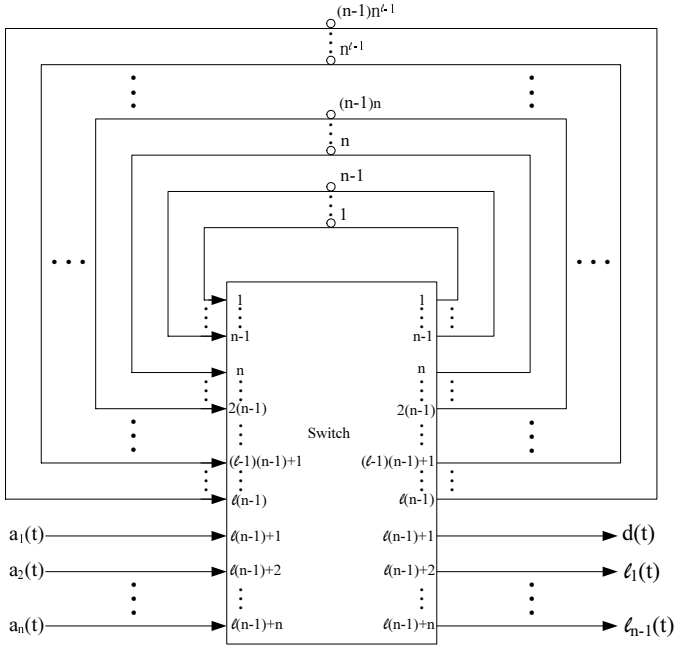

Fig. 5. A construction of a self-routing optical n-to-1 FIFO multiplexer with buffer size $n^\ell - 1$ by using an optical $((n-1)\ell + n) \times ((n-1)\ell + n)$ (bufferless) crossbar switch and $(n-1)\ell$ fiber delay lines.

In our constructions of optical priority queues in Figure 3(a), we have used optical nFM1's. Recall that an optical nFM1 with buffer size B is a concatenation of an optical nFM with buffer size $B-1$ and a fiber delay line with delay equal to one as shown in Figure 3(b). It was shown in [4, Figure 3] and [19, Lemma 2] that a self-routing optical nFM with buffer size $n^\ell - 1$ can be constructed by using a feedback system consisting of an optical $((n-1)\ell + n) \times ((n-1)\ell + n)$ (bufferless) crossbar switch and $(n-1)\ell$ fiber delay lines (see Figure 5). As such, an optical nFM with buffer size $B-1$ can be constructed by using the feedback system in Figure 5 with $\ell = \lceil \log_n B \rceil$. It then follows that an optical nFM1 with buffer size B can be constructed by using an optical $((n-1)\lceil \log_n B \rceil + n + 1) \times ((n-1)\lceil \log_n B \rceil + n + 1)$ (bufferless) crossbar switch and $(n-1)\lceil \log_n B \rceil + 1$ fiber delay lines.

Suppose $1 \leq s \leq k-1$ and $m \geq 1$. We see from Theorem 7 and the argument in the above paragraph that if $n, B_1, B_2, \ldots, B_k$, $|\Psi_1|, |\Psi_2|, \ldots, |\Psi_k|$ satisfy the conditions (A1)–(A3),

then an optical priority queue with buffer size $U_k = \sum_{i=1}^{k} |\Psi_i|$ can be constructed by using an optical $(M+2) \times (M+2)$ (bufferless) crossbar switch and M fiber delay lines, where $M = m \sum_{i=1}^{k} ((n-1)\lceil \log_n B_i \rceil + n + 1)$. Apparently, the construction complexity of such a construction is increasing with n (as the switch size $M+2$ is increasing with n). To achieve minimum construction complexity, we have to choose n as small as possible, and it is clear from (A1) that we should choose $n = \min\{2s+1, k\} + 1$.

Furthermore, when the achieved buffer size is used as the performance measure of a construction, we have to choose $|\Psi_i|$ as large as possible for $i = 1, 2, \ldots, k$ to maximize the achieved buffer size $U_k = \sum_{i=1}^{k} |\Psi_i|$ in our construction. From (A3), it is clear that we should choose $|\Psi_i| = (m-1)B_i + 1$ for $i = 1, 2, \ldots, k$. If $m = 1$, then $|\Psi_i| = 1$ for $i = 1, 2, \ldots, k$, and hence we have $U_k = k$. In this case, we should choose $B_i = 1$ for $i = 1, 2, \ldots, k$ (as there is at most $|\Psi_i| = 1$ packet routed to the i^{th} group of nFM1's at any slot according to the routing policy (R3)), and hence we have $M = k(n+1)$. Clearly this is not an interesting case as $U_k = M/(n+1)$ grows only linearly with M. So we assume that $m \geq 2$ in the rest of this paper.

For $m \geq 2$, the choice $|\Psi_i| = (m-1)B_i + 1$ can be made as large as possible by choosing B_i as large as possible for $i = 1, 2, \ldots, k$. From (A2), we can see that we should make the following choice:

(A2*) If $s+1 \leq k \leq 2s+2$, then

$$B_i = B_{k-i+1} = \begin{cases} 1, \text{ if } i = 1, \\ \sum_{j=1}^{i-1} ((m-1)B_j + 1), \text{ if } 2 \leq i \leq \lceil k/2 \rceil. \end{cases}$$

On the other hand, if $k \geq 2s+3$, then

$$B_i = B_{k-i+1} = \begin{cases} 1, \text{ if } i = 1, \\ \sum_{j=1}^{i-1} ((m-1)B_j + 1), \text{ if } 2 \leq i \leq s+1, \\ \sum_{j=i-s}^{i-1} ((m-1)B_j + 1), \text{ if } s+2 \leq i \leq \lceil k/2 \rceil. \end{cases}$$

We summarize the above findings in the following theorem.

Theorem 9 *Suppose $1 \leq s \leq k-1$ and $m \geq 2$. Then an optical priority queue with buffer size U_k can be constructed by using a feedback system consisting of an optical $(M+2) \times (M+2)$ (bufferless) crossbar switch and M fiber delay lines, where*

$$U_k = \sum_{i=1}^{k} ((m-1)B_i + 1), \tag{22}$$

$$M = m \sum_{i=1}^{k} ((n-1)\lceil \log_n B_i \rceil + n + 1), \tag{23}$$

in which $n = \min\{2s+1, k\} + 1$ and $B_1, B_2, \ldots, B_k$ are given by (A2*).

To express the buffer size U_k (given by (22)) in terms of M (given by (23)), we need the following results on the buffer sizes $B_1, B_2, \ldots, B_k$ given by (A2*).

Theorem 10 *Suppose that $1 \le s \le k-1$, $m \ge 2$, and $B_1, B_2, \ldots, B_k$ are given by (A2*).*

(i) If $s = 1$, then we have

$$B_i = B_{k-i+1} = \sum_{j=0}^{i-1} (m-1)^j = \begin{cases} i, \text{ if } m = 2 \text{ and } 1 \le i \le \lceil k/2 \rceil, \\ \frac{(m-1)^i - 1}{m-2}, \text{ if } m \ge 3 \text{ and } 1 \le i \le \lceil k/2 \rceil. \end{cases} \tag{24}$$

(ii) If $s \ge 2$ and $s+1 \le k \le 2s+2$, then we have

$$B_i = B_{k-i+1} = \begin{cases} 1, \text{ if } i = 1, \\ m^{i-1} + \frac{m^{i-2}-1}{m-1}, \text{ if } 2 \le i \le \lceil k/2 \rceil. \end{cases} \tag{25}$$

(iii) If $s \ge 2$ and $k \ge 2s+3$, then we have

$$B_i = B_{k-i+1} = \sum_{j=1}^{s} \alpha_j \lambda_j^i - \frac{s}{s(m-1)-1} \text{ for } 2 \le i \le \lceil k/2 \rceil, \tag{26}$$

where $\lambda_1, \lambda_2, \ldots, \lambda_s$ are the roots of the characteristic polynomial $p(z) = z^s - \sum_{j=0}^{s-1}(m-1)z^j$ associated with the s^{th}-order nonhomogeneous linear difference equation with constant coefficients given by $B_i = \sum_{j=i-s}^{i-1}((m-1)B_j + 1)$ for $s+2 \le i \le \lceil k/2 \rceil$, and $\alpha_1, \alpha_2, \ldots, \alpha_s$ can be obtained by solving the s equations $B_i = m^{i-1} + \frac{m^{i-2}-1}{m-1}$, $i = 2, 3, \ldots, s+1$.

We need the following two lemmas (whose proofs are given in Appendix A and Appendix B, respectively) for the proof of Theorem 10.

Lemma 11 *Suppose that $m \ge 2$ and assume that $x_1 = 1$ and $x_i = \sum_{j=1}^{i-1}((m-1)x_j + 1)$ for $i \ge 2$.*

(i) $x_i = mx_{i-1} + 1$ for $i \ge 3$.

(ii) $x_i = m^{i-1} + \frac{m^{i-2}-1}{m-1}$ for $i \ge 2$.

Lemma 12 *Suppose that $s \ge 2$ and $m \ge 2$, and suppose that $p(z)$ is a polynomial in the indeterminate z given by $p(z) = z^s - \sum_{j=0}^{s-1}(m-1)z^j$.*

(i) $p(z)$ has s complex roots and they are all distinct.

(ii) If s is odd, then $p(z)$ has one positive root and $s-1$ nonreal roots. On the other hand, if s is even, then $p(z)$ has one positive root, one negative root, and $s-2$ nonreal roots.

(iii) The positive root, say λ_+, of $p(z)$ lies in the open interval $(m - 1/(m-1)^{s-1}, m - (m-1)/m^s)$, and $\lambda_+ \approx m$ for sufficiently large s or m. The roots of $p(z)$ other than λ_+ lie in the annulus $\{z \in C : \lambda_+/(\lambda_+ + 1) \le |z| \le \lambda_+ - m + 1\}$. Therefore, λ_+ is the root of $p(z)$ of the largest magnitude.

Proof. (Proof of Theorem 10) As we have from (A2*) that $B_i = B_{k-i+1}$ for $1 \le i \le \lceil k/2 \rceil$, it suffices to prove the theorem for B_i for $1 \le i \le \lceil k/2 \rceil$.

(i) Suppose $s = 1$. Then we have $k \geq s + 1 = 2$. First consider the case that $k = 2$. In this case, we have from (A2*) that $B_1 = B_2 = 1$ and hence (24) holds.

Now consider the case that $k \geq 3$. We will prove by induction on i that $B_i = \sum_{j=0}^{i-1}(m-1)^j$ for $1 \leq i \leq \lceil k/2 \rceil$. It is clear from (A2*) that $B_1 = 1$. Assume as the induction hypothesis that $B_{i-1} = \sum_{j=0}^{i-2}(m-1)^j$ for some $1 \leq i-1 \leq \lceil k/2 \rceil - 1$. Then we have from (A2*) (note that $s = 1$ in this case) and the induction hypothesis that

$$B_i = (m-1)B_{i-1} + 1 = (m-1)\sum_{j=0}^{i-2}(m-1)^j + 1 = \sum_{j=0}^{i-1}(m-1)^j.$$

(ii) Suppose $s \geq 2$ and $s + 1 \leq k \leq 2s + 2$. Then we have from (A2*) that $B_1 = 1$ and $B_i = \sum_{j=1}^{i-1}((m-1)B_j + 1)$ for $2 \leq i \leq \lceil k/2 \rceil$, and hence (25) follows immediately from Lemma 11(ii).

(iii) Suppose $s \geq 2$ and $k \geq 2s + 3$. Then we have from (A2*) that $B_1 = 1$ and $B_i = \sum_{j=1}^{i-1}((m-1)B_j + 1)$ for $2 \leq i \leq s + 1$, and hence it follows from Lemma 11(ii) that $B_i = m^{i-1} + \frac{m^{i-2}-1}{m-1}$ for $2 \leq i \leq s + 1$.

From (A2*), we also have the s^{th}-order nonhomogeneous linear difference equation with constant coefficients given by $B_i = \sum_{j=i-s}^{i-1}((m-1)B_j+1)$ for $s+2 \leq i \leq \lceil k/2 \rceil$. The characteristic polynomial $p(z)$ associated with this difference equation is given by $p(z) = z^s - \sum_{j=0}^{s-1}(m-1)z^j$. As we know from Lemma 12(i) that $p(z)$ has s roots, say $\lambda_1, \lambda_2, \ldots, \lambda_s$, and they are all distinct, it then follows from well-established results in the theory of difference equations [28, Chapter 2] that $B_i = \sum_{j=1}^{s} \alpha_j \lambda_j^i + \alpha_0$ for $2 \leq i \leq \lceil k/2 \rceil$, where $\alpha_0 = -\frac{s}{s(m-1)-1}$ is a particular solution to this difference equation, and $\alpha_1, \alpha_2, \ldots \alpha_s$ can be obtained by solving the s equations $B_i = m^{i-1} + \frac{m^{i-2}-1}{m-1}$, $i = 2, 3, \ldots, s + 1$. ■

In the following theorem (whose proof is given in Appendix C), we use the results in Theorem 10 to express the buffer size U_k (given by (22)) in terms of M (given by (23)).

Theorem 13 *Suppose that $1 \leq s \leq k-1$, $m \geq 2$, U_k is given by (22), and M is given by (23).*

(i) If $k = 2$, then we have $U_k = M/4$.

(ii) If $s = 1$, $k \geq 3$, and $m = 2$, then we have $M/7 \leq U_k \leq (M/13)^2$.

(iii) If $s = 1$, $k \geq 3$, and $m \geq 3$, then we have

$$2^{\sqrt{2M\log_2(m-1)/(3m)} - 6\log_2(m-1)} \leq U_k \leq 2^{\sqrt{2M\log_2(m-1)/(3m)} + \log_2(8(m-1))}. \tag{27}$$

Therefore, we have $U_k = 2^{O(\sqrt{2M\log_2(m-1)/(3m)})}$ in this case.

(iv) If $s \geq 2$, $s + 1 \leq k \leq 2s$, and $m \geq 2$, then we have

$$2^{\sqrt{M\log_2(k+1)\log_2 m/(km)} - 4\log_2(k+1)} \leq U_k \leq 2^{\sqrt{M\log_2(k+1)\log_2 m/(km)} + \log_2(3m)}. \tag{28}$$

Therefore, we have $U_k = 2^{O(\sqrt{M\log_2(k+1)\log_2 m/(km)})}$ in this case.

(v) If $s \geq 2$, $k \geq 2s+1$, $m \geq 2$, and if we approximate B_i and B_{k-i+1} by $B_i = B_{k-i+1} \approx \alpha_+ m^i$ for $1 \leq i \leq \lceil k/2 \rceil$, where α_+ is the coefficient of the term λ_+^i in (26) with λ_+ being the positive root of the characteristic polynomial $p(z)$ in Theorem 10(iii), then we have

$$U_k \approx 2^{\sqrt{M \log_2 (2s+2) \log_2 m/((2s+1)m)} + \log_2(\alpha_+ m)}. \tag{29}$$

Therefore, we have $U_k \approx 2^{O(\sqrt{M \log_2 (2s+2) \log_2 m/((2s+1)m)})}$ in this case.

Remark 14 *(i) The reason for the approximation $B_i = B_{k-i+1} \approx \alpha_+ m^i$ for $1 \leq i \leq \lceil k/2 \rceil$ in Theorem 13(v) is as follows. In this case, the roots $\lambda_1, \lambda_2, \ldots, \lambda_s$ of the polynomial $p(z)$ and the coefficients $\alpha_1, \alpha_2, \ldots, \alpha_s$ in the expression for $B_1, B_2, \ldots, B_k$ in (26) cannot be easily expressed in terms of s, k, and m. However, for this case we know from Lemma 12(iii) that $\lambda_+ > m - 1/(m-1)^{s-1} \geq 1$ and any root λ of $p(z)$ other than λ_+ has magnitude $|\lambda| \leq \lambda_+ - m + 1 < 1 - (m-1)/m^s < 1$. Thus, for sufficiently large s and m, we can approximate B_i and B_{k-i+1} by only keeping the term $\alpha_+ \lambda_+^i$ in (26). For sufficiently large s and m, we also know from Lemma 12(iii) that $\lambda_+ \approx m$ and hence we can approximate B_i and B_{k-i+1} by $B_i = B_{k-i+1} \approx \alpha_+ m^i$.*

(ii) From Theorem 13, we see that we can achieve a buffer size U_k that goes beyond polynomial in M when $s = 1$, $k \geq 3$, and $m \geq 3$, or when $s \geq 2$, $k \geq 3$, and $m \geq 2$. In these cases, we can achieve a buffer size of $U_k = 2^{O(\sqrt{\alpha M})}$, where α is a constant that depends on s, k, and m.

(iii) For $s = 1$ and $k \geq 3$, we see from Theorem 13 (iii) that among the integers $m \geq 3$, it is better to choose $m = 5$ (as $(1/m)\log_2(m-1)$ achieves its maximum when $m = 5$), and hence we can achieve a buffer size of $U_k = 2^{O(\sqrt{4M/15})}$.

(iv) From Theorem 13(iv), we see that among the integers $m \geq 2$, it is better to choose $m = 3$ (as $(1/m)\log_2 m$ achieves its maximum when $m = 3$), and hence we can achieve a buffer size of $U_k = 2^{O(\sqrt{(\log_2 3)(\log_2(k+1))M/(3k)})}$. If we need to achieve a larger buffer size, then we need to choose a larger k. For $k \geq 6$, the result $U_k = 2^{O(\sqrt{(\log_2 3)(\log_2(k+1))M/(3k)})}$ is worse than $U_k = 2^{O(\sqrt{4M/15})}$ in Remark 14(iii). A similar remark can be made by using the result in Theorem 13(v).

V. Router Buffer Sizing, Numerical Results, and Feasibility Issues

A. Router Buffer Sizing

The buffer sizes in today's commercial backbone routers are in the order of millions of packets. This follows from the well-known *rule of thumb* (or the *bandwidth-delay product (BDP) rule*) $B = C \times RTT$, where B is the buffer size of the router, C is the data rate of the bottleneck link, and RTT is the average round-trip time of flows passing through the bottleneck link. This rule was proposed by Villamizar and Song in 1994 [29] in order to guarantee 100% link utilization.

When the bottleneck link carries a large number, say N, of desynchronized long-lived TCP flows, researchers from Stanford University appealed to statistical multiplexing and claimed in 2004 [30] that the buffer size follows the *small-buffer rule* (or the *Stanford model*) $B =$

$C \times RTT/\sqrt{N}$, and the buffer size can be dramatically reduced to hundreds or thousands of packets while achieving near 100% link utilization at the same time. When the traffic comes from slower access networks, or when the source paces the packets it sends, it was claimed in [31] that the buffer size can be further reduced to tens of packets, but at the expense of sacrificing about 15% of link capacity. This is known as the *tiny-buffer rule* in the literature. However, it was mentioned in [32] that the small-buffer or tiny-buffer rule may not hold when there is a small number of flows, or when there is a very skewed mix of short-lived and long-lived flows. Whether the small/tiny-buffer rules hold for most parts of today's backbone networks remains an open issue worthy of further investigation [32]. We note that similar findings from different perspectives are presented in [33] and [34]. We refer to [35] for a comprehensive review of the buffer sizing problem.

As mentioned in Section I, one of the primary technological bottlenecks in all-optical packet switching is the difficulty in building *large* optical buffers. Since optical buffers are very costly but optical link capacity is abundant in today's optical technology, all-optical packet switching currently is most feasible in the tiny-buffer regime by trading off capacity for tiny buffers. However, with the advances in optical technology, it is possible to build all-optical packet-switched networks for all regimes of buffer sizes in a cost-effective manner in the future.

B. Numerical Results

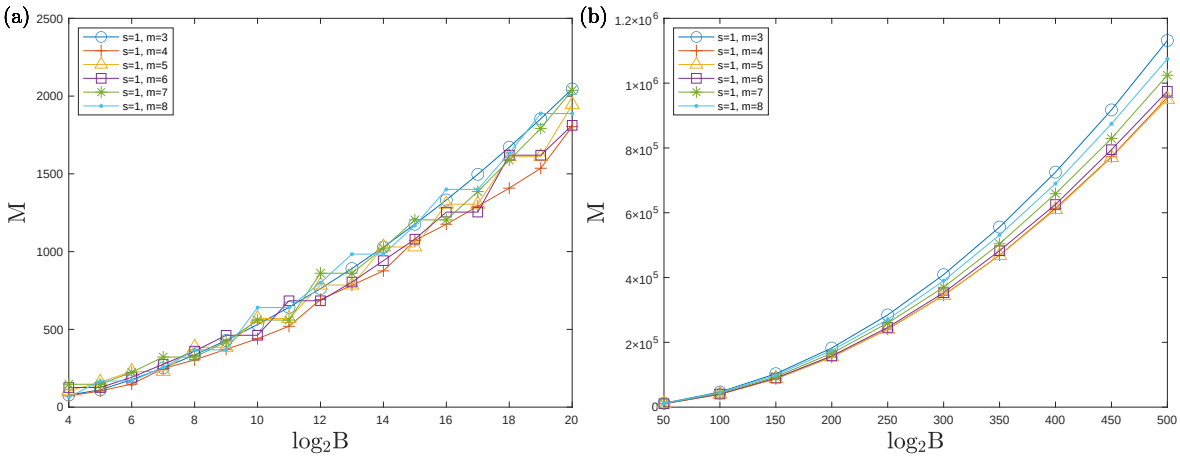

Fig. 6. The value of M required to achieve a targeted buffer size B for $s = 1$ and $3 \leq m \leq 8$. (a) $2^4 \leq B \leq 2^{20}$. (b) $2^{50} \leq B \leq 2^{500}$.

We have shown in Remark 14(ii) that by using a feedback system consisting of an optical $(M+2) \times (M+2)$ (bufferless) crossbar switch and M fiber delay lines, we can achieve a buffer size U_k of $2^{O(\sqrt{\alpha M})}$ when $s = 1$, $k \geq 3$, and $m \geq 3$, or when $s \geq 2$, $k \geq 3$, and $m \geq 2$, where U_k is given by (22), M is given by (23), and α is a constant that depends on s, k, and m. It is known [36] that an $N \times N$ switch can be built by using $N \log_2 N - N/2$ 2×2 switches via the Benes network. Thus, the construction complexity in our constructions is increasing with M.

For a targeted buffer size B and for given values of s and m, we first use (22) to choose the smallest k such that $U_k \geq B$ in order to achieve minimum construction complexity (as M is increasing with k), and then use (23) to calculate the value of M required to achieve the targeted buffer size B. In Figure 6, we show the results for $s = 1$ and $3 \leq m \leq 8$. It is clear from Figure 6 that M is roughly proportional to $(\log_2 B)^2$, which conforms with the result that $B \approx U_k = 2^{O(\sqrt{\alpha M})}$ in Remark 14(ii). For $2^4 \leq B \leq 2^{20}$, we see from Figure 6(a) that the best choice of m is $m = 4$ as it achieves minimum construction complexity. As B gets larger, say $2^{50} \leq B \leq 2^{500}$, we see from Figure 6(b) that the best choice of m is $m = 5$ as expected from our theoretical analysis in Remark 14(iii). We note that similar results hold for $s \geq 2$. In particular, for $s = 2$ and $s = 3$, we find that the best choice of m is $m = 5$ for $2^4 \leq B \leq 2^{20}$.

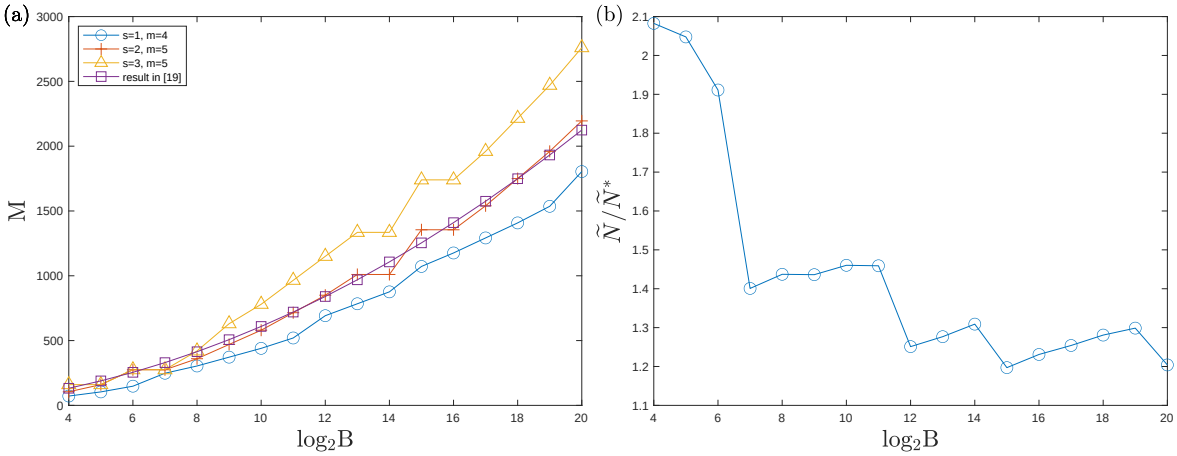

Fig. 7. (a) The value of M required to achieve a targeted buffer size B for $s = 1$ and $m = 4$, for $s = 2$ and $m = 5$, for $s = 3$ and $m = 5$, and for the construction in [19]. (b) The ratio between the numbers of 2×2 switches required to achieve a targeted buffer size B for the constructions in [19] and for our constructions with $s = 1$ and $m = 4$.

The construction in [19] is a special case of our constructions with $s = 1$, $k = 2\ell - 1$ for some $\ell \geq 2$, $m = 3$, $n = 4$, $B_1 = B_{2\ell-1} = 1$, $B_i = B_{2\ell-i} = 2^{i-2}$ for $2 \leq i \leq \ell$, and $|\Psi_i| = |\Psi_{2\ell-i}| = 2^{i-1}$ for $1 \leq i \leq \ell$. From $U_k = \sum_{i=1}^{k} |\Psi_i|$ and (23), we obtain

$$U_k = 3 \cdot 2^{\ell-1} - 2 \text{ and } M = (9\ell^2 + 33\ell - 12)/2. \tag{30}$$

In Figure 7(a), we show the value of M required to achieve the targeted buffer size B for $s = 1$ and $m = 4$, for $s = 2$ and $m = 5$, for $s = 3$ and $m = 5$, and for the construction in [19] by using (30). From Figure 7(a), we see that the construction complexity is increasing with s, and hence we should choose $s = 1$ so as to achieve minimum construction complexity. Furthermore, our constructions with $s = 1$ and $m = 4$ has lower construction complex than that of the construction in [19] for $2^4 \leq B \leq 2^{20}$.

Note that we have mentioned above that an $(M + 2) \times (M + 2)$ switch can be built by using $(M + 2) \log_2 (M + 2) - (M + 2)/2$ 2×2 switches. Let $\widetilde{N}$ and $\widetilde{N}^*$ be the numbers of 2×2 switches required to achieve a targeted buffer size B for the construction in [19] and for our

constructions with $s = 1$ and $m = 4$, respectively. We see from Figure 7(b) that $\widetilde{N}/\widetilde{N}^*$ is between 1.2 and 2.08 for $2^4 \leq B \leq 2^{20}$. This means that the cost of the construction in [19] is 1.2 to 2.08 times of that of our constructions with $s = 1$ and $m = 4$. Indeed, the actual saving by using our constructions could be significant as an optical 2×2 switch is still quite expensive currently. This can be seen from Table I that the extra number $\widetilde{N} - \widetilde{N}^*$ of 2×2 switches needed is quite large and ranges from 457 to 4634 for $2^4 \leq B \leq 2^{20}$. Even in the tiny/small buffer regime, say, $2^4 \leq B \leq 2^{13}$, the extra number of 2×2 switches is still quite large and ranges from 457 to 1983.

Furthermore, we also see from Table I that the number of 2×2 switches needed for our constructions with $s = 1$ and $m = 4$ goes from 423 for $B = 2^4$ in the tiny-buffer regime to 1867 for $B = 2^7$ in the small-buffer regime, and to 8147 for $B = 2^{14}$ beyond the small-buffer regime. This means that the cost for larger buffers in the small-buffer regime (resp., beyond the small-buffer regime) is about 5 times (resp., 20 times) higher than that in the tiny-buffer regime.

B	2^4	2^5	2^6	2^7	2^8	2^9	2^{10}	2^{11}	2^{12}
$\widetilde{N}$	880	1352	1929	2615	3412	4323	5350	6496	7762
$\widetilde{N}^*$	423	661	1010	1867	2374	3010	3664	4452	6204
$\widetilde{N} - \widetilde{N}^*$	457	691	919	748	1038	1313	1686	2044	1558
$\widetilde{N}/\widetilde{N}^*$	2.0804	2.0454	1.9099	1.4006	1.4372	1.4362	1.4602	1.4591	1.2511
B	2^{13}	2^{14}	2^{15}	2^{16}	2^{17}	2^{18}	2^{19}	2^{20}	
$\widetilde{N}$	9151	10664	12303	14069	15964	17990	20148	22438	
$\widetilde{N}^*$	7168	8147	10277	11430	12730	14046	15514	18636	
$\widetilde{N} - \widetilde{N}^*$	1983	2517	2026	2639	3234	3944	4634	3802	
$\widetilde{N}/\widetilde{N}^*$	1.2766	1.3089	1.1971	1.2309	1.2540	1.2808	1.2987	1.2040	

TABLE I
THE NUMBERS OF 2×2 SWITCHES REQUIRED TO ACHIEVE A TARGETED BUFFER SIZE B FOR THE CONSTRUCTION IN [19] AND FOR OUR CONSTRUCTIONS WITH $s = 1$ AND $m = 4$.

C. Feasibility Issues

In [31], an optical buffer was demonstrated by using semiconductor optical amplifier (SOA) gate matrix 2×2 switches and waveguide delay lines. The SOA gate matrix switches exhibit high extinction ratios (>40 dB), low crosstalk (<-40 dB), and fast switching times (1-ns rise time 20%–80%), which results in longer storage times, lower crosstalk interference, and higher throughput. The waveguide delay lines have low loss (on the order of 0.01 dB/cm) and can be integrated into a small size. Such an implementation of optical buffers has been demonstrated to be a viable approach for high-speed buffering of hundreds of packets.

In reality, crosstalk due to power leakage from other optical links, power loss experienced during recirculations through the optical switches and the fiber delay lines, amplified spontaneous emission (ASE) from the Erbium-doped fiber amplifiers (EDFA) that are used for boosting the signal power, and the pattern effect of the optical switches, among others, lead to a limitation on

the number of times that an optical packet can recirculate through the optical switches and the fiber delay lines. This is because an optical packet recirculating more than a limited number of times cannot be reliably recognized at the destined output port due to severe power loss and/or serious noise accumulation even if it appears at the right place and at the right time.

With the technological advances in 3R (reamplification, retiming, and reshaping) regeneration, hundreds of packet recirculations is possible [31]. Thus, the issue of limited number of packet recirculations may somewhat be alleviated. In the scenario that the numbers of recirculations of most packets are below certain threshold beyond which the received packets cannot be recognized, the issue of limited number of packet recirculations may not be a serious problem. For example, in [37] we have considered the SDL constructions of optical priority queues in [15] under i.i.d. Bernoulli arrival traffic, i.i.d. Bernoulli control input, and uniform priority assignment. When the arrival rate, say 0.9, is less than the departure request rate, say 0.95, our analytical results and simulation results show that the average number of packet recirculations is less than two and the probability that a packet recirculating more than 15 times is less than 10^{-4} (see [37, Figure 3]). This shows that the limited number of packet recirculations may not be a serious issue in this case.

Ideally, it would be much better to have a systematic approach that is technology-independent to build optical buffers with a limited number of packet recirculations. Results along this line on the constructions of optical 2FM's with a limited number of packet recirculations can be found in [8] and [9].

Another important practical issue of concern is fault-tolerant capability. In the design of a network element, survivability deals with the situation that some of the components of the network element may not function properly. Without taking the survivability aspect into consideration during the design process, a network element consisting of hundreds or thousands of components may be in a total breakdown even when only a single component fails to function properly. As before, it would be nice to have a technology-independent approach to build optical buffers with fault-tolerant capability. Results along this line on the constructions of fault-tolerant optical 2FM's and fault-tolerant optical linear compressors/decompressors can be found in [7] and [38].

VI. Conclusion

In this paper, we have shown that the feedback system in Figure 3(a) can be operated as an optical priority queue under a simple priority-based routing policy. The idea is to first route the packet with the highest (resp., lowest) priority to the departure (resp., loss) link whenever necessary, and then route the other packets at the input links of the crossbar switch to the optical nFM1's according to their buffering tags. We have also shown that by using a feedback system consisting of an optical $(M+2)\times(M+2)$ (bufferless) crossbar switch and M fiber delay lines, we can achieve a buffer size of $2^{O(\sqrt{\alpha M})}$, where α is a constant that depends on s, k, and m. Furthermore, we showed that the best buffer size that we can achieve is $2^{O(\sqrt{4M/15})}$. Our result (exponential in $\sqrt{M}$) substantially improves on the best known result (polynomial in M) in the

literature. From our numerical results, we showed that the best choice of s is $s = 1$, and it is best to choose $s = 1$ and $m = 4$ in the tiny/small-buffer regime. We also showed that the construction complexity of our constructions is lower than that of the construction in [19], and the actual saving, in terms of the number of 2×2 switches needed, by our constructions could be quite significant even in the tiny-buffer and small-buffer regimes. Finally, we note that there is still a gap between the buffer sizes of our constructions and the theoretical upper bound 2^M. Whether this theoretical upper bound can be achieved or not and, in the case that it can be achieved, how to achieve it remains a very challenging open problem.

APPENDIX A
PROOF OF LEMMA 11

(i) Suppose that $i \geq 3$. Then we have

$$\begin{aligned} x_i &= \sum_{j=1}^{i-1}((m-1)x_j + 1) = \sum_{j=1}^{i-2}((m-1)x_j + 1) + (m-1)x_{i-1} + 1 \\ &= x_{i-1} + (m-1)x_{i-1} + 1 = mx_{i-1} + 1. \end{aligned}$$

(ii) We prove (ii) by induction on i. Clearly, we have $x_2 = (m-1)x_1+1 = (m-1)\cdot 1+1 = m$. Assume as the induction hypothesis that $x_{i-1} = m^{i-2} + \frac{m^{i-3}-1}{m-1}$ for some $i - 1 \geq 2$. Then we have from (i) (note that $i \geq 3$) and the induction hypothesis that

$$x_i = mx_{i-1} + 1 = m\left(m^{i-2} + \frac{m^{i-3}-1}{m-1}\right) + 1 = m^{i-1} + \frac{m^{i-2}-1}{m-1}.$$

APPENDIX B
PROOF OF LEMMA 12

(i) Since $p(z)$ is clearly a polynomial of degree s, it follows from the fundamental theorem of algebra [20, Theorem 16.22] that $p(z)$ has s complex roots.

Now we show that the roots of $p(z)$ are distinct by contradiction. Assume on the contrary that $p(z)$ has a repeated root, say λ. Let $f(z)$ be a polynomial given by $f(z) = (z-1)p(z)$. Then we have

$$f(z) = (z-1)p(z) = (z-1)\left(z^s - \sum_{j=0}^{s-1}(m-1)z^j\right) = z^{s+1} - mz^s + m - 1. \tag{31}$$

As it is clear that λ is also a repeated root of $f(z)$, we must have $f'(\lambda) = 0$. Thus, we see from (31) that $(s+1)\lambda^s - m \cdot s\lambda^{s-1} = 0$, and it follows from $s \geq 2$ that either $\lambda = 0$ or $\lambda = \frac{m\cdot s}{s+1}$. Since 0 cannot be a root of $p(z)$ (as $p(0) = -(m-1) \neq 0$), we must have

$$\lambda = \frac{m \cdot s}{s+1}. \tag{32}$$

From $\lambda = \frac{m \cdot s}{s+1} \neq 1$ (as $(m-1)s \neq 1$), we see that

$$\begin{aligned} p(\lambda) &= \lambda^s - \sum_{j=0}^{s-1}(m-1)\lambda^j = \lambda^s - \frac{(m-1)(\lambda^s-1)}{\lambda-1} \\ &= \lambda^s - \frac{(m-1)(s+1)(\lambda^s-1)}{(m-1)s-1} = \frac{(m-1)(s+1)-m\lambda^s}{(m-1)s-1}, \end{aligned}$$

and it then follows from $p(\lambda) = 0$ that

$$\lambda^s = \frac{(m-1)(s+1)}{m}. \tag{33}$$

If $m = 2$, then we see from $\lambda = \frac{m \cdot s}{s+1} = \frac{2s}{s+1}$ in (32) and $s \geq 2$ that

$$\begin{aligned} \lambda^s - \frac{(m-1)(s+1)}{m} &= \left(\frac{2s}{s+1}\right)^s - \frac{s+1}{2} = \left(1 + \frac{s-1}{s+1}\right)^s - \frac{s+1}{2} \\ &\geq 1 + \binom{s}{1}\frac{s-1}{s+1} - \frac{s+1}{2} = \frac{(s-1)^2}{2(s+1)} > 0, \end{aligned}$$

and we have reached a contradiction to (33) in this case. On the other hand, if $m \geq 3$, then we see from $\lambda = \frac{m \cdot s}{s+1}$ in (32) and $s \geq 2$ that

$$\lambda^s = \left(\frac{m \cdot s}{s+1}\right)^s \geq \left(\frac{3 \cdot 2}{2+1}\right)^s = 2^s > s+1 > \frac{(m-1)(s+1)}{m},$$

and we have also reached a contradiction to (33) in this case.

(ii) To prove (ii), we need Descartes' rule of signs [21]–[24], which says that the number $z_+(f)$ of positive roots (counting multiplicities) of a nonzero polynomial $f(z)$ with real coefficients is at most equal to the number $v(f)$ of changes of signs in the sequence of the polynomial's coefficients (omitting the zero coefficients), and that the difference between these two numbers is even, i.e.,

$$v(f) - z_+(f) \text{ is a nonnegative even integer.} \tag{34}$$

It is easy to see from (34) and $z_+(f) \geq 0$ that

$$\text{if } v(f) \leq 1, \text{ then } z_+(f) = v(f). \tag{35}$$

Note that as it is clear that $v(p) = 1$, we have from (35) that $z_+(p) = v(p) = 1$, i.e., $p(z)$ has exactly one positive root. Also note that 0 cannot be a root of $p(z)$ (as $p(0) = -(m-1) \neq 0$).

Let $f(z) = (z-1)p(z)$ and let $g(z) = f(-z) = -(z+1)p(-z)$. Then it is easy to see that a positive number λ is a root of $g(z)$ if and only if $-\lambda$ is a root of $p(z)$. We first consider the case that s is odd. In this case , we see from (31) that $g(z) = z^{s+1} + mz^s + m - 1$. As it is clear that $v(g) = 0$ in this case, we have from (35) that $z_+(g) = v(g) = 0$, i.e., $g(z)$ has no positive roots, or, equivalently, $p(z)$ has no negative roots. Therefore, we conclude from (i) and the results above that $p(z)$ has one positive root and $s-1$ nonreal roots in this case.

Now we consider the case that s is even. In this case, we see from (31) that $g(z) = -z^{s+1} - mz^s + m - 1$. As it is clear that $v(g) = 1$ in this case, we have from (35) that $z_+(g) = v(g) = 1$, i.e., $g(z)$ has exactly one positive root, or, equivalently, $p(z)$ has exactly one negative root. Therefore, we conclude from (i) and the results above that $p(z)$ has one positive root, one negative root, and $s-2$ nonreal roots in this case.

(iii) As $p(m-1) = -\sum_{j=0}^{s-2}(m-1)^{j+1} < 0$ (note that $s \geq 2$) and $p(m) = 1 > 0$, we have from the intermediate-value theorem for continuous functions [20, Theorem 4.33] that $m - 1 < \lambda_+ < m$. From (31), we see that

$$\lambda_+^{s+1} - m\lambda_+^s + m - 1 = (\lambda_+ - 1)p(\lambda_+) = (\lambda_+ - 1) \cdot 0 = 0,$$

and it then follows that

$$\lambda_+ = m - \frac{m-1}{\lambda_+^s}. \tag{36}$$

Since $s \geq 2$, $m \geq 2$, and we have proved that $m - 1 < \lambda_+ < m$, we deduce from (36) that $m - 1/(m-1)^{s-1} < \lambda_+ < m - (m-1)/m^s$ and hence $\lambda_+ \approx m$ for sufficiently large s or m.

To show that the roots of $p(z)$ other than λ_+ lie in the annulus $\{z \in C : \lambda_+/(\lambda_+ + 1) \leq |z| \leq \lambda_+ - m + 1\}$, we need Eneström-Kakeya theorem [25]–[27, Theorem 4], which says that if $f(z) = \sum_{i=0}^{\ell} b_i z^i$ is a polynomial of degree ℓ with positive coefficients, i.e., $b_i > 0$ for $i = 0, 1, \ldots, \ell$, then all the roots of $f(z)$ lie in the annulus $\{z \in C : \min_{1 \leq i \leq \ell} \frac{b_{i-1}}{b_i} \leq |z| \leq \max_{1 \leq i \leq \ell} \frac{b_{i-1}}{b_i}\}$.

Since λ_+ is a root of $p(z)$, we can write $p(z)$ as $p(z) = (z - \lambda_+)q(z)$ for some polynomial $q(z) = \sum_{i=0}^{s-1} b_i z^i$. By comparing the coefficients of the term z^i in $(z - \lambda_+)q(z)$ and $p(z)$ for $i = 0, 1, \ldots, s$, it is easy to see that $-\lambda_+ b_0 = -m+1$, $b_{i-1} - \lambda_+ b_i = -m+1$ for $i = 1, 2, \ldots, s-1$, and $b_{s-1} = 1$. In the following, we prove by induction on i that

$$b_i = \sum_{j=0}^{i} \frac{m-1}{\lambda_+^{j+1}} \text{ for } i = 0, 1, \ldots, s-1. \tag{37}$$

Note that from $-\lambda_+ b_0 = -m+1$, we have $b_0 = (m-1)/\lambda_+$. Assume as the induction hypothesis that $b_{i-1} = \sum_{j=0}^{i-1} \frac{m-1}{\lambda_+^{j+1}}$ for some $0 \leq i - 1 \leq s - 2$. Then we have from $b_{i-1} - \lambda_+ b_i = -m + 1$ and the induction hypothesis that

$$b_i = \frac{1}{\lambda_+}(b_{i-1} + m - 1) = \frac{1}{\lambda_+} \sum_{j=0}^{i-1} \frac{m-1}{\lambda_+^{j+1}} + \frac{m-1}{\lambda_+} = \sum_{j=0}^{i} \frac{m-1}{\lambda_+^{j+1}}.$$

From $b_{i-1} - \lambda_+ b_i = -m + 1$ for $1 \leq i \leq s-1$, $0 < b_0 < b_1 < \cdots < b_{s-1}$ (by (37)), $b_1 = \frac{m-1}{\lambda_+^2}(\lambda_+ + 1)$ in (37), and $b_{s-1} = 1$, we obtain

$$\min_{1 \leq i \leq s-1} \frac{b_{i-1}}{b_i} = \min_{1 \leq i \leq s-1} \left(\lambda_+ - \frac{m-1}{b_i} \right) = \lambda_+ - \frac{m-1}{b_1} = \frac{\lambda_+}{\lambda_+ + 1} \tag{38}$$

and

$$\max_{1\le i\le s-1}\frac{b_{i-1}}{b_i}=\max_{1\le i\le s-1}\left(\lambda_+-\frac{m-1}{b_i}\right)=\lambda_+-\frac{m-1}{b_{s-1}}=\lambda_+-m+1. \tag{39}$$

Therefore, we have from Eneström-Kakeya theorem, (38), and (39) that all the roots of $q(z)$, i.e., all the roots of $p(z)$ other than λ_+, lie in the annulus $\{z\in C:\lambda_+/(\lambda_++1)\le|z|\le\lambda_+-m+1\}$.

APPENDIX C
PROOF OF THEOREM 13

(i) Suppose that $k=2$. Then we have $s=1$ (as $1\le s\le k-1=1$) and $n=\min\{2s+1,k\}+1=k+1=3$. We also have from (A2*) that $B_1=B_2=1$. As such, it follows from (22) that $U_k=\sum_{i=1}^{2}((m-1)\cdot 1+1)=2m$ and it follows from (23) that $M=m\sum_{i=1}^{2}(2\lceil\log_3 1\rceil+4)=8m$. Thus, we have $U_k=M/4$.

(ii) Suppose that $s=1$, $k\ge 3$, and $m=2$. Then we have $n=\min\{2s+1,k\}+1=2s+2=4$ and we have from (24) that $B_i=B_{k-i+1}=i$ for $1\le i\le\lceil k/2\rceil$. If $k=3$, then we have $B_1=1$, $B_2=2$, and $B_3=1$, and it follows from (22) that $U_k=2+3+2=7$ and from (23) that $M=2(5+8+5)=36$. Thus, we have $M/7\le U_k\le(M/13)^2$ in this case. So we assume that $k\ge 4$ in the rest of the proof. We consider the following two cases.

Case 1: k is odd, say $k=2\ell-1$ for some $\ell\ge 3$. It follows from (22) that

$$U_k=2\sum_{i=1}^{\ell-1}(i+1)+(\ell+1)=\ell^2+2\ell-1, \tag{40}$$

and it follows from (23) that

$$\begin{aligned}M&=2\left[2\sum_{i=1}^{\ell-1}(3\lceil\log_4 i\rceil+5)+(3\lceil\log_4\ell\rceil+5)\right]\\&=12\left[\sum_{j=1}^{\ell'}j(4^j-4^{j-1})+(\ell'+1)(\ell-1-4^{\ell'})\right]+6\lceil\log_4\ell\rceil+10(2\ell-1)\\&=4(3\ell'+8)\ell-12\ell'-4^{\ell'+2}+6\lceil\log_4\ell\rceil-18,\end{aligned} \tag{41}$$

where ℓ' is the unique nonnegative integer such that $4^{\ell'}+1\le\ell-1\le 4^{\ell'+1}$. From (41), $0\le\ell'\le\log_4(\ell-2)\le(\ell-3)/2$ (note that $\ell\ge 3$), and (40), we have

$$\begin{aligned}M&\le 4(3\ell'+8)\ell-4(\ell-1)+6(\ell'+2)-18\le 6(\ell-3)\ell+28\ell+3(\ell-3)-2\\&=6\ell^2+13\ell-11\le 7U_k,\end{aligned} \tag{42}$$

$$M\ge 4(3\ell'+8)\ell-12\ell'-16(\ell-2)+6(\ell'+1)-18\ge 16\ell+20\ge 13\sqrt{U_k}. \tag{43}$$

Thus, we see from (42) and (43) that $M/7\le U_k\le(M/13)^2$.

Case 2: k is even, say $k=2\ell$ for some $\ell\ge 2$. It follows from (22) that

$$U_k=2\sum_{i=1}^{\ell}(i+1)=\ell^2+3\ell, \tag{44}$$

and it follows from (23) that

$$M = 2 \cdot 2 \sum_{i=1}^{\ell} (3\lceil \log_4 i \rceil + 5) = 12 \left[\sum_{j=1}^{\ell'} j(4^j - 4^{j-1}) + (\ell' + 1)(\ell - 4^{\ell'}) \right] + 20\ell$$
$$= 4(3\ell' + 8)\ell - 4^{\ell'+2} + 4, \tag{45}$$

where ℓ' is the unique nonnegative integer such that $4^{\ell'} + 1 \le \ell \le 4^{\ell'+1}$. From (45), $0 \le \ell' \le \log_4(\ell - 1) \le (\ell - 2)/2$, and (44), we have

$$M \le 4(3\ell' + 8)\ell - 4\ell + 4 \le 6(\ell - 2)\ell + 28\ell + 4 = 6\ell^2 + 16\ell + 4 \le 7U_k, \tag{46}$$
$$M \ge 4(3\ell' + 8)\ell - 16(\ell - 1) + 4 \ge 16\ell + 20 \ge 13\sqrt{U_k}. \tag{47}$$

Thus, we see from (46) and (47) that $M/7 \le U_k \le (M/13)^2$.

(iii) Suppose that $s = 1$, $k \ge 3$, and $m \ge 3$. Then we have $n = \min\{2s+1, k\}+1 = 2s+2 = 4$ and we have from (24) that $B_i = B_{k-i+1} = ((m-1)^i - 1)/(m-2)$ for $1 \le i \le \lceil k/2 \rceil$. We consider the following two cases.

Case 1: k is odd, say $k = 2\ell - 1$ for some $\ell \ge 2$. It follows from (22) that

$$U_k = 2 \sum_{i=1}^{\ell-1} [(m-1)((m-1)^i - 1)/(m-2) + 1] + (m-1)((m-1)^\ell - 1)/(m-2) + 1$$
$$= [m(m-1)^{\ell+1} - 2(m-1)^2 - (2\ell - 1)(m-2)]/(m-2)^2. \tag{48}$$

From (48), $\ell \ge 2$, and $m \ge 3$, we see that

$$U_k \ge [m(m-1)^{\ell+1} - 2(m-1)^\ell - (2\ell-1)(m-2)]/(m-2)^2$$
$$= (m-1)^\ell + (3(m-1)^\ell - 2\ell + 1)/(m-2) \ge (m-1)^\ell \tag{49}$$
$$U_k \le m(m-1)^{\ell+1}/(m-2)^2 \le 8(m-1)^\ell. \tag{50}$$

For $1 \le i \le \ell$, it is easy to see that $(m-1)^{i-1} \le B_i \le (m-1)^i$, and hence we have $\lceil \log_4 B_i \rceil < \log_4 B_i + 1 \le i \cdot \log_4(m-1) + 1$ and $\lceil \log_4 B_i \rceil \ge \log_4 B_i \ge (i-1)\log_4(m-1)$. As such, it follows from (23) that

$$\frac{M}{m} \le 2 \sum_{i=1}^{\ell-1} [3(i \cdot \log_4(m-1) + 1) + 5] + 3(\ell \cdot \log_4(m-1) + 1) + 5$$
$$= 3\ell^2 \log_4(m-1) + 8(2\ell - 1) \le 3(\ell + 6)^2 \log_4(m-1), \tag{51}$$
$$\frac{M}{m} \ge 2 \sum_{i=1}^{\ell-1} [3(i-1)\log_4(m-1) + 5] + 3(\ell - 1)\log_4(m-1) + 5$$
$$\ge 3(\ell-1)^2 \log_4(m-1) + 5(2\ell - 1) \ge 3(\ell-1)^2 \log_4(m-1). \tag{52}$$

Thus, (27) follows from (49), $\ell \ge \sqrt{2M/(3m\log_2(m-1))} - 6$ in (51), (50), and $\ell \le \sqrt{2M/(3m\log_2(m-1))} + 1$ in (52).

Case 2: k is even, say $k = 2\ell$ for some $\ell \geq 2$. It follows from (22) that

$$\begin{aligned} U_k &= 2\sum_{i=1}^{\ell}[(m-1)((m-1)^i - 1)/(m-2) + 1] \\ &= [2(m-1)^{\ell+2} - 2(m-1)^2 - 2\ell(m-2)]/(m-2)^2. \end{aligned} \tag{53}$$

From (53), $\ell \geq 2$, and $m \geq 3$, we see that

$$\begin{aligned} U_k &\geq [2(m-1)^{\ell+2} - 2(m-1)^{\ell} - 2\ell(m-2)]/(m-2)^2 \\ &= 2(m-1)^{\ell} + (4(m-1)^{\ell} - 2\ell)/(m-2) \geq (m-1)^{\ell}, \end{aligned} \tag{54}$$

$$U_k \leq 2(m-1)^{\ell+2}/(m-2)^2 \leq 8(m-1)^{\ell}. \tag{55}$$

As in Case 1 above, it follows from (23) that

$$\begin{aligned} \frac{M}{m} &\leq 2\sum_{i=1}^{\ell}[3(i \cdot \log_4(m-1) + 1) + 5] = 3\ell(\ell+1)\log_4(m-1) + 16\ell \\ &\leq 3(\ell+6)^2 \log_4(m-1), \end{aligned} \tag{56}$$

$$\begin{aligned} \frac{M}{m} &\geq 2\sum_{i=1}^{\ell}[3(i-1)\log_4(m-1) + 5] = 3\ell(\ell-1)\log_4(m-1) + 10\ell \\ &\geq 3(\ell-1)^2 \log_4(m-1). \end{aligned} \tag{57}$$

Thus, (27) also follows from (54)–(57) as in Case 1 above.

(iv) Suppose that $s \geq 2$, $s+1 \leq k \leq 2s$, and $m \geq 2$. Then we have $n = \min\{2s+1, k\} + 1 = k+1$ and we have from (25) that $B_1 = B_k = 1$ and $B_i = B_{k-i+1} = m^{i-1} + (m^{i-2} - 1)/(m-1)$ for $2 \leq i \leq \lceil k/2 \rceil$. We consider the following two cases.

Case 1: k is odd, say $k = 2\ell - 1$ for some $\lceil s/2 \rceil + 1 \leq \ell \leq s$. It follows from (22) that

$$\begin{aligned} U_k &= 2m + 2\sum_{i=2}^{\ell-1}[(m-1)(m^{i-1} + (m^{i-2} - 1)/(m-1)) + 1] \\ &\quad + (m-1)(m^{\ell-1} + (m^{\ell-2} - 1)/(m-1)) + 1 \\ &= m^{\ell} + ((m^2+1)m^{\ell-2} - 2)/(m-1). \end{aligned} \tag{58}$$

From (58), $\ell \geq \lceil s/2 \rceil + 1 \geq 2$, and $m \geq 2$, we see that

$$m^{\ell} \leq U_k \leq 3m^{\ell}. \tag{59}$$

For $1 \le i \le \ell$, it is easy to see that $m^{i-1} \le B_i \le (k+1)m^{i-1}$, and hence we have $\lceil \log_{k+1} B_i \rceil < \log_{k+1} B_i + 1 \le (i-1)\log_{k+1} m + 2$ and $\lceil \log_{k+1} B_i \rceil \ge \log_{k+1} B_i \ge (i-1)\log_{k+1} m$. As such, it follows from (23) that

$$\begin{aligned}
\frac{M}{m} &\le 2\sum_{i=1}^{\ell-1}[k((i-1)\log_{k+1} m + 2) + k + 2] + k((\ell-1)\log_{k+1} m + 2) + k + 2 \\
&= k(\ell-1)^2 \log_{k+1} m + (3k+2)(2\ell-1) \le k(\ell + 4/\log_{k+1} m)^2 \log_{k+1} m \\
&= k(\ell + 4\log_2(k+1)/\log_2 m)^2 \log_2 m/\log_2(k+1), \qquad (60) \\
\frac{M}{m} &\ge 2\sum_{i=1}^{\ell-1}[k(i-1)\log_{k+1} m + k + 2] + k(\ell-1)\log_{k+1} m + k + 2 \\
&= k(\ell-1)^2 \log_{k+1} m + (k+2)(2\ell-1) \ge k(\ell-1)^2 \log_2 m/\log_2(k+1). \qquad (61)
\end{aligned}$$

Thus, (28) follows from (59), $\ell \ge \sqrt{M\log_2(k+1)/(km\log_2 m)} - 4\log_2(k+1)/\log_2 m$ in (60), and $\ell \le \sqrt{M\log_2(k+1)/(km\log_2 m)} + 1$ in (61).

Case 2: k is even, say $k = 2\ell$ for some $\lceil (s+1)/2 \rceil \le \ell \le s$. It follows from (22) that

$$\begin{aligned}
U_k &= 2m + 2\sum_{i=2}^{\ell}[(m-1)(m^{i-1} + (m^{i-2}-1)/(m-1)) + 1] \\
&= 2m^\ell + 2(m^{\ell-1}-1)/(m-1). \qquad (62)
\end{aligned}$$

From (62), $\ell \ge \lceil (s+1)/2 \rceil \ge 2$, and $m \ge 2$, we see that

$$m^\ell \le U_k \le 3m^\ell. \qquad (63)$$

As in Case 1 above, it follows from (23) that

$$\begin{aligned}
\frac{M}{m} &\le 2\sum_{i=1}^{\ell}[k((i-1)\log_{k+1} m + 2) + k + 2] \\
&= k\ell(\ell-1)\log_{k+1} m + (3k+2)\cdot 2\ell \le k(\ell + 4/\log_{k+1} m)^2 \log_{k+1} m \\
&= k(\ell + 4\log_2(k+1)/\log_2 m)^2 \log_2 m/\log_2(k+1), \qquad (64) \\
\frac{M}{m} &\ge 2\sum_{i=1}^{\ell}[k(i-1)\log_{k+1} m + k + 2] = k\ell(\ell-1)\log_{k+1} m + (k+2)\cdot 2\ell \\
&\ge k(\ell-1)^2 \log_2 m/\log_2(k+1). \qquad (65)
\end{aligned}$$

Thus, (28) also follows from (63)–(65) as in Case 1 above.

(v) Suppose that $s \ge 2$, $k \ge 2s+1$, and $m \ge 2$. Then we have $n = \min\{2s+1, k\} + 1 = 2s+2$. Assume that $B_i = B_{k-i+1} \approx \alpha_+ m^i$ for $1 \le i \le \lceil k/2 \rceil$. We consider the following two cases.

Case 1: k is odd, say $k = 2\ell - 1$ for some $\ell \geq s + 1$. It follows from (22) that

$$\begin{aligned} U_k &\approx 2\sum_{i=1}^{\ell-1}[(m-1)\alpha_+ m^i + 1] + (m-1)\alpha_+ m^\ell + 1 \\ &= 2\alpha_+(m^\ell - m) + \alpha_+(m-1)m^\ell + 2\ell - 1 \\ &\approx \alpha_+ m^{\ell+1}, \end{aligned} \tag{66}$$

and it follows from (23) that

$$\begin{aligned} \frac{M}{m} &\approx 2\sum_{i=1}^{\ell-1}[(2s+1)\log_{2s+2}(\alpha_+ m^i) + 2s + 3] + (2s+1)\log_{2s+2}(\alpha_+ m^\ell) + 2s + 3 \\ &= (2s+1)\ell^2 \log_{2s+2} m + ((2s+1)\log_{2s+2}\alpha_+ + 2s + 3)(2\ell - 1) \\ &\approx (2s+1)\ell^2 \log_2 m / \log_2(2s+2). \end{aligned} \tag{67}$$

Thus, (29) follows from (66) and $\ell \approx \sqrt{M \log_2(2s+2)/((2s+1)m\log_2 m)}$ in (67).

Case 2: k is even, say $k = 2\ell$ for some $\ell \geq s + 1$. It follows from (22) that

$$U_k = 2\sum_{i=1}^{\ell}[(m-1)\alpha_+ m^i + 1] = 2\alpha_+(m^{\ell+1} - m) + 2\ell \approx \alpha_+ m^{\ell+1}, \tag{68}$$

and it follows from (23) that

$$\begin{aligned} \frac{M}{m} &\approx 2\sum_{i=1}^{\ell}[(2s+1)\log_{2s+2}(\alpha_+ m^i) + 2s + 3] \\ &= (2s+1)\ell(\ell+1)\log_{2s+2} m + 2\ell((2s+1)\log_{2s+2}\alpha_+ + 2s + 3) \\ &\approx (2s+1)\ell^2 \log_2 m / \log_2(2s+2). \end{aligned} \tag{69}$$

Thus, (29) follows from (68) and (69) as in Case 1 above.

REFERENCES

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

[2] Y.-T. Chen, C.-S. Chang, J. Cheng, and D.-S. Lee, "Feedforward SDL constructions of output-buffered multiplexers and switches with variable length bursts," in *Proceedings IEEE International Conference on Computer Communications (INFOCOM'07)*, Anchorage, AK, USA, May 6–12, 2007.

[3] I. Chlamtac, A. Fumagalli, and C.-J. Suh, "Optimal 2×1 multi-stage optical packet multiplexer," in *Proceedings IEEE Global Telecommunications Conference (GLOBECOM'97)*, Phoenix, AZ, USA, November 3–8, 1997, pp. 566–570.

[4] C.-S. Chang, D.-S. Lee, and C.-K. Tu, "Recursive construction of FIFO optical multiplexers with switched delay lines," *IEEE Transactions on Information Theory*, vol. 50, pp. 3221–3233, December 2004.

[5] C.-S. Chang, D.-S. Lee, and C.-K. Tu, "Using switched delay lines for exact emulation of FIFO multiplexers with variable length bursts," *IEEE Journal on Selected Areas in Communications*, vol. 24, pp. 108–117, April 2006.

[6] C.-C. Chou, C.-S. Chang, D.-S. Lee and J. Cheng, "A necessary and sufficient condition for the construction of 2-to-1 optical FIFO multiplexers by a single crossbar switch and fiber delay lines," *IEEE Transactions on Information Theory*, vol. 52, pp. 4519–4531, October 2006.

[7] J. Cheng, "Constructions of fault tolerant optical 2-to-1 FIFO multiplexers," *IEEE Transactions on Information Theory*, vol. 53, pp. 4092–4105, November 2007.

[8] J. Cheng, "Constructions of optical 2-to-1 FIFO multiplexers with a limited number of recirculations," *IEEE Transactions on Information Theory*, vol. 54, pp. 4040–4052, September 2008.

[9] J. Cheng, C.-S. Chang, S.-H. Yang, T.-H. Chao, D.-S. Lee, and C.-M. Lien, "Greedy constructions of optical queues with a limited number of recirculations," *IEEE Transactions on Information Theory*, vol. 63, pp. 5314–5326, August 2017. Conference version appeared in *IEEE INFOCOM* 2008.

[10] X. Wang, X. Jiang, and S. Horiguchi, "Improved bounds on the feedfoward design of optical multiplexers," in *Proceedings International Symposium on Parallel Architectures, Algorithms, and Networks (I-SPAN'08)*, Sydney, Australia, May 7–9, 2008, pp. 178–183.

[11] X. Wang, X. Jiang, and A. Pattavina, "Constructing N-to-N shared optical queues with switches and fiber delay lines," *IEEE Transactions on Information Theory*, vol. 58, pp. 3836–3842, June 2012.

[12] S.-Y. R. Li and X. J. Tan, "Mux/demux queues, FIFO queues, and their construction by fiber memories," *IEEE Transactions on Information Theory*, vol. 57, pp. 1328–1343, March 2011. Conference version appeared in *Allerton* 2006.

[13] A. D. Sarwate and V. Anantharam, "Exact emulation of a priority queue with a switch and delay lines," *Queueing Systems: Theory and Applications*, vol. 53, pp. 115–125, July 2006.

[14] H.-C. Chiu, C.-S. Chang, J. Cheng, and D.-S. Lee, "A simple proof for the constructions of optical priority queues," *Queueing Systems: Theory and Applications*, vol. 56, pp. 73–77, June 2007.

[15] H.-C. Chiu, C.-S. Chang, J. Cheng, and D.-S. Lee, "Using a single switch with $O(M)$ inputs/outputs for the construction of an optical priority queue with $O(M^3)$ buffer," in *Proceedings IEEE International Conference on Computer Communications (INFOCOM'07 Minisymposium)*, Anchorage, AK, USA, May 6–12, 2007.

[16] H. Kogan and I. Keslassy, "Optimal-complexity optical router," in *Proceedings IEEE International Conference on Computer Communications (INFOCOM'07 Minisymposium)*, Anchorage, AK, USA, May 6–12, 2007.

[17] H. Rastegarfar, M. Ghobadi, and Y. Ganjali, "Emulation of Optical PIFO Buffers," in *Proceedings IEEE Global Communications Conference (GLOBECOM'09)*, Honolulu, HI, USA, November 30–December 4, 2009.

[18] A. Datta, "Construction of polynomial-size optical priority queues using linear switches and fiber delay lines," *IEEE/ACM Transactions on Networking*, vol. 25, pp. 974–987, April 2017.

[19] B. Tang, X. Wang, C.-T. Nguyen, B. Ye, and S. Lu, "Construction of subexponential-size optical priority queues with switches and fiber delay lines," *IEEE/ACM Transactions on Networking*, vol. 28, pp. 336–346, February 2020.

[20] T. M. Apostol, *Mathematical Analysis*, 2nd ed. Reading, MA: Addison-Wesley, 1974.

[21] R. Descartes, *La Géométrie (Discours de la Méthode, third part)*, ed. of Leiden, 1637, p. 373.

[22] R. Descartes, *The Geometry of René Descartes with a Facsimile of the First Edition (trans. D. E. Smith and M. L. Latham)*, Mineola, NY: Dover, 1954.

[23] D. J. Struik, Ed., *A Source Book in Mathematics 1200–1800*, Princeton, NJ: Princeton University Press, 1986, pp. 89–93.

[24] X. Wang, "A simple proof of Descartes's rule of signs," *American Mathematical Monthly*, vol. 111, pp. 525–526, June–July 2004.

[25] G. Eneström, "Härledning af en allmän formel för antalet pensionärer, som vid en godtycklig tidpunkt förefinnas inom en sluten pensionskassa," *Öfversigt af Kungl. Vetenskap-Akademiens Förhandlinger*, vol. 50, pp. 405–415, 1893.

[26] S. Kakeya, "On the limits of the roots of an algebraic equation with positive coefficients," *Tòhoku Mathematical Journal*, vol. 2, pp. 140–142, 1912.

[27] R. B. Gardner and N. K. Govil, "Eneström-Kakeya theorem and some of its generalizations," in *Current Topics in Pure and Computational Complex Analysis*, S. Joshi, M. Dorff, and I. Lahiri, Eds., ser. Trends in Mathematics. New Delhi, India: Birkhäuser/Springer, 2014, ch. 8, pp. 171–199.

[28] S. Elaydi, *An Introduction to Difference Equations*, 3rd ed. New York, NY: Springer, 2005.

[29] C. Villamizar and C. Song, "High performance TCP in ANSNET," *ACM SIGCOMM Computer Communication Review*, vol. 24, pp. 45–60, October 1994.

[30] G. Appenzeller, I. Keslassy, and N. McKeown, "Sizing router buffers," *ACM SIGCOMM Computer Communication Review*, vol. 34, pp. 281–292, October 2004.

[31] N. Beheshti, E. Burmeister, Y. Ganjali, J. E. Bowers, D. J. Blumenthal, and N. McKeown, "Optical packet buffers for backbone Internet routers," *IEEE/ACM Transactions on Networking*, vol. 18, pp. 1599–1609, October 2010.

[32] N. McKeown, G. Appenzeller, and I. Keslassy, "Sizing router buffers (Redux)," *ACM SIGCOMM Computer Communication Review*, vol. 49, pp. 69–74, October 2019.

[33] R. S. Prasad, C. Dovrolis, and M. Thottan, "Router buffer sizing for TCP traffic and the role of the output/input capacity ratio," *IEEE/ACM Transactions on Networking*, vol. 17, pp. 1645–1658, October 2009.

[34] A. Lakshmikantha, C. Beck, and R. Srikant, "Impact of file arrivals and departures on buffer sizing in core routers," *IEEE/ACM Transactions on Networking*, vol. 19, pp. 347–358, April 2011.

[35] A. Vishwanath, V. Sivaraman, and M. Thottan, "Perspectives on router buffer sizing: Recent results and open problems," *ACM SIGCOMM Computer Communication Review*, vol. 39, pp. 34–39, April 2009.

[36] C.-S. Chang and D.-S. Lee, *Principles, Architectures and Mathematical Theories of High Performance Switches*, Hsinchu, Taiwan: National Tsing Hua University Press, 2008.

[37] J. Cheng, X.-C. Huang, C.-H. Cheng, H.-H. Chou, C.-S. Chang, and D.-S. Lee, "Average number of recirculations in SDL constructions of optical priority queues," *IEEE Communications Letters*, vol. 15, pp. 899–901, August 2011.

[38] C.-S. Chang, J. Cheng, T.-H. Chao, and D.-S. Lee, "Optimal constructions of fault tolerant optical linear compressors and linear decompressors," *IEEE Transactions on Communications*, vol. 57, pp. 1140–1150, April 2009. Conference version appeared in *IEEE INFOCOM* 2007.