Load Balanced Birkhoff-von Neumann Switches, Part II: Multi-stage Buffering

Cheng-Shang Chang, Duan-Shin Lee, and Ching-Ming Lien [1]

Institute of Communications Engineering
National Tsing Hua University
Hsinchu 300, Taiwan, R.O.C.

Abstract

The main objective of this sequel is to solve the out-of-sequence problem that occurs in the load balanced Birkhoff-von Neumann switch with one-stage buffering. We do this by adding a load-balancing buffer in front of the first stage and a resequencing-and-output buffer after the second stage. Moreover, packets are distributed at the first stage according to their *flows*, instead of their *arrival times* in Part I. In this paper, we consider multicasting flows with two types of scheduling policies: the First Come First Served (FCFS) policy and the Earliest Deadline First (EDF) policy. The FCFS policy requires a jitter control mechanism in front of the second stage to ensure proper ordering of the traffic entering the second stage. For the EDF scheme, there is no need for jitter control. It uses the departure times of the corresponding FCFS output-buffered switch as deadlines and schedules packets according to their deadlines. For both policies, we show that the end-to-end delay through our multi-stage switch is bounded above by the sum of the delay from the corresponding FCFS output-buffered switch and a constant that only depends on the size of the switch and the number of multicasting flows supported by the switch.

Key words: multi-stage switches, load balancing, scheduling, multicasting, performance bounds

Email addresses: `cschang@ee.nthu.edu.tw`, `lds@cs.nthu.edu.tw`, `keiichi@gibbs.ee.nthu.edu.tw` (Cheng-Shang Chang, Duan-Shin Lee, and Ching-Ming Lien).

[1] This research is supported in part by the National Science Council, Taiwan, R.O.C., under Contracts NSC-89-2213-E007-127, NSC-89-2219-E007-015 and the program for promoting academic excellence of universities 89-E-FA04-1-4.

1 Introduction

In Part I [5], we proposed the load balanced Birkhoff-von Neumann switch with one-stage buffering (see Figure 1). Such a switch consists of two stages of crossbar switching fabrics and one stage of buffering. The buffer at the input port of the second stage uses the Virtual Output Queueing (VOQ) technique to solve the problem of head-of-line blocking. In such a switch, packets are of the same size. Also, time is slotted and synchronized so that exactly one packet can be transmitted within a time slot. In a time slot, both crossbar switches set up connection patterns corresponding to permutation matrices that are periodically generated from a one-cycle permutation matrix.

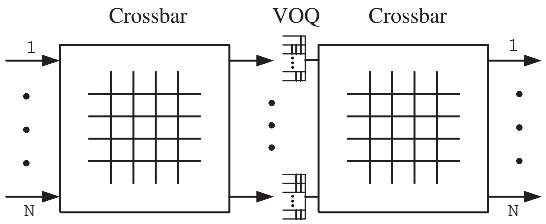

Fig. 1. The load balanced Birkhoff-von Neumann switch with one-stage buffering

The reasoning behind such a switch architecture is as follows: since the connection patterns are periodic, packets from the same input port of the first stage are distributed in a round-robin fashion to the second stage according to their arrival times. Thus, the first stage performs load balancing for the incoming traffic. As the traffic coming into the second stage is load balanced, it suffices to use simple periodic connection patterns to perform switching at the second stage. This is shown in [5] as a special case of the original Birkhoff [1]-von Neumann [13] decomposition used in [3,4]. There are several advantages of using such an architecture, including scalability, low hardware complexity, 100% throughput, low average delay in heavy load and bursty traffic, and efficient buffer usage. However, the main drawback of the load balanced Birkhoff-von Neumann switch with one-stage buffering is that packets might be out of sequence.

The main objective of this sequel is to solve the out-of-sequence problem that occurs in the load balanced Birkhoff-von Neumann switch with one-stage buffering. One quick fix is to add a resequencing-and-output buffer after the second stage. However, as packets are distributed according to their *arrival times* at the first stage, there is no guarantee on the size of the resequencing-and-output buffer to prevent packet losses. For this, one needs to distributed packets according to their *flows*, as indicated in the paper by Iyer and McKeown [9]. This is done by adding a flow splitter and a load-balancing buffer in front of the first stage (see Figure 2). For an $N \times N$ switch, the load-balancing buffer at each input port of the first stage consists of N virtual output queues

(VOQ) destined for the N output ports of that stage. Packets from the same *flow* are split in the round-robin fashion to the N virtual output queues and scheduled under the First Come First Served (FCFS) policy. By so doing, load balancing can be achieved for each flow as packets from the same flow are split almost evenly to the input ports of the second stage. More importantly, as pointed out in [9], the delay and the buffer size of the load-balancing buffer are bounded by constants that only depend on the size of the switch and the number of flows. The resequencing-and-output buffer after the second stage not only performs resequencing to keep packets in sequence, but also stores packets waiting for transmission from the output links.

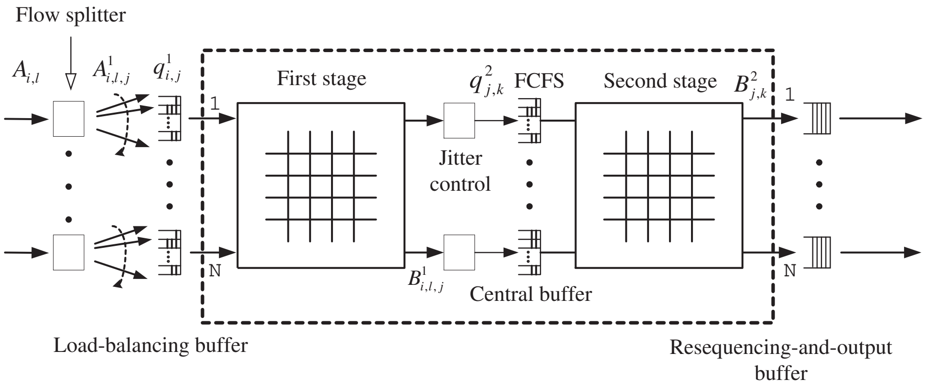

Fig. 2. The load balanced switch with multi-stage buffering under FCFS

In this paper, we consider a traffic model with multicasting flows. This is a more general model than the point-to-point traffic model in [9]. A multicasting flow is a stream of packets that has one common input and a set of common outputs. For the multicasting flows, fan-out splitting (see e.g., [8]) is performed at the central buffers (the VOQ in front of the second stage). The central buffers are assumed to be infinite so that no packets are lost in the switch. We consider two types of scheduling policies in the central buffers: the FCFS policy and the Earliest Deadline First (EDF) policy. For the FCFS policy, a jitter control mechanism is added in the VOQ in front of the second stage. Such a jitter control mechanism delays every packet to its maximum delay at the first stage so that the flows entering the second stage are simply time-shifted flows of the original ones. Our main result for the FCFS scheme with jitter control is the following theorem. The proof of Theorem 1 will be given in Section 2.

Theorem 1 *Suppose that all the buffers are empty at time 0. Then the following results hold for FCFS scheme with jitter control.*

(i) *The end-to-end delay for a packet through our switch with multi-stage buffering is bounded above by the sum of the delay through the corresponding FCFS output-buffered switch and* $(N-1)L_{\max} + NM_{\max}$*, where* $L_{\max}$ *(resp.* $M_{\max}$*) is the maximum number of flows at an input (resp. output) port.*

(ii) *The load-balancing buffer at an input port of the first stage is bounded above by $NL_{\max}$.*
(iii) *The delay through the load-balancing buffer at an input port of the first stage is bounded above by $(N-1)L_{\max}$.*
(iv) *The resequencing-and-output buffer at an output port of the second stage is bounded above $NM_{\max}$.*

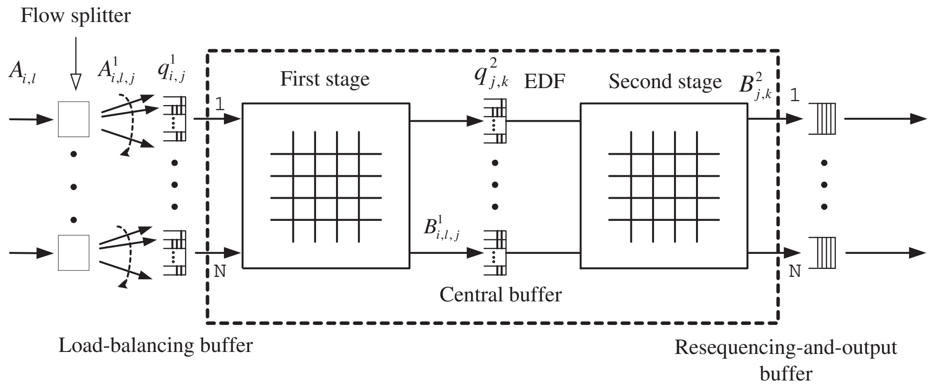

Fig. 3. The load balanced switch with multi-stage buffering under EDF

In the EDF scheme (see Figure 3), every packet is assigned a deadline that is the departure time from the corresponding FCFS output-buffered switch. Packets are scheduled according to their deadlines in the central buffers. For the EDF scheme, there is no need to implement the jitter control mechanism in the FCFS scheme. As such, average packet delay can be greatly reduced. However, as there is no jitter control, one might need a larger resequencing buffer than that in the FCFS scheme with jitter control. Since the first stage is the same as that in the FCFS scheme, the delay and the buffer size are still bounded by $(N-1)L_{\max}$ and $NL_{\max}$ respectively. Moreover, we show the following theorem for the EDF scheme. Its proof will be given in Section 3.

Theorem 2 *Suppose that all the buffers are empty at time 0. Then the following results hold for the EDF scheme.*

(i) *The end-to-end delay for a packet through our switch with multi-stage buffering is bounded above by the sum of the delay through the corresponding FCFS output-buffered switch and $(N-1)(L_{\max}+M_{\max})$.*
(ii) *The resequencing-and-output buffer at an output port of the second stage is bounded above $(N-1)(L_{\max}+M_{\max})$.*

Computing the departure times from the corresponding FCFS output-buffered switch needs global information of all the inputs. A simple way is to use the packet arrival times as deadlines. Then the EDF scheme based on arrival times yields the same departure order except those packets that arrives at same time. Since there are at most $M_{\max}$ packets that can arrive at the same time to an output port of the corresponding output-buffered switch, the end-to-end delay for a packet through the multi-stage switch using arrival times as deadlines is bounded above by the sum of the delay through the corresponding FCFS output-buffered switch and $(N-1)L_{\max}+NM_{\max}$. Also, the resequencing-and-output buffer at an output port of the second stage in this case is bounded above $(N-1)L_{\max}+NM_{\max}$.

It is of some interest to compare our schemes with the Combined Input-Output Queueing (CIOQ) switch (see e.g., [6,12]). The CIOQ switch provides the exact emulation of the corresponding output-buffered switch. For our schemes, we only have bounded differences between the departure times in our schemes and the departure times of the corresponding output-buffered switch. However, the CIOQ switch requires an internal rate speedup of two and a complicated scheduling algorithm that prohibits its practical use. In our schemes, there is no need for internal rate speedup (it is done by two switching fabrics) and the complexity of the scheduling algorithm is low ($O(1)$ for the FCFS scheme).

2 The FCFS scheme

In this section, we prove Theorem 1. As discussed in the Introduction, we consider $N \times N$ switches with multicasting flows under the FCFS scheduling policy. To be precise, let L_i be the number of multicasting flows through the i^{th} input port. Denote by $A_{i,\ell}(t)$ the cumulative number of packet arrivals by time t from the ℓ^{th} multicasting flow at the i^{th} input port, $i=1,\dots,N$, $\ell=1,\dots,L_i$. Also, let $S_{i,\ell}$ be the set of outputs of that flow, $S^*(k)=\{(i,\ell): k \in S_{i,\ell}\}$ be the set of multicasting flows through the k^{th} output, and $M_k = |S^*(k)|$ be the number of multicasting flows through the k^{th} output port. Define $L_{\max} = \max_{1\le i\le N} L_i$ as the maximum number of multicasting flow through an input port and $M_{\max} = \max_{1\le k\le N} M_k$ as the the maximum number of multicasting flow through an output port.

2.1 Analysis for the output-buffered switch under FCFS

Now consider feeding these multicasting flows to an $N \times N$ output-buffered switch under the FCFS policy (see Figure 4). Assume that there is an infinite buffer at each output port and that all the buffers are empty at time 0. Let $A_k^o(t)$ be the cumulative number of arrivals at the k^{th} output buffer by time t, $q_k^o(t)$ be the number of packets at the k^{th} output buffer at time t, and $B_k^o(t)$ be the cumulative number of departures at the k^{th} output buffer by time t. Since an output-buffered switch is a work-conserving link with the constant capacity

1, one has the following well-known representation (see e.g., [2], Lemma 1.3.1)

$$q_k^o(t) = \max_{0 \le s \le t}[A_k^o(t) - A_k^o(s) - (t-s)], \tag{1}$$

where

$$A_k^o(t) = \sum_{(i,\ell) \in S^*(k)} A_{i,\ell}(t). \tag{2}$$

Moreover,

$$B_k^o(t) = A_k^o(t) - q_k^o(t) = \min_{0 \le s \le t}[A_k^o(s) + (t-s)]. \tag{3}$$

As the scheduling policy is FCFS and there are at most M_k arrivals per unit of time, packets that arrive at the k^{th} buffer at time t will depart between $t + q_k^o(t) - M_k + 1$ and $t + q_k^o(t)$. Also, note from (3) that

$$B_k^o(t) - B_k^o(s) \le t - s, \quad \text{for all } s \le t, \tag{4}$$

as there is at most one packet coming out from an output port per time slot.

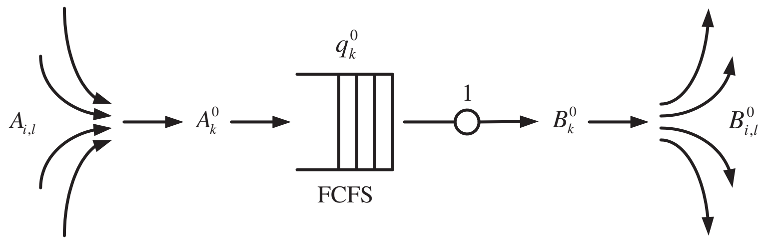

Fig. 4. The FCFS output-buffered switch with multicasting flows

2.2 Analysis for the load-balancing buffer

As discussed in the Introduction, the load-balancing buffer at each input port of the first stage consists of N virtual output queues (VOQ) destined for the N output ports of that stage. Packets from the same flow are split in the round-robin fashion to the N virtual output queues and scheduled under the FCFS policy. Without loss of generality, we assume that the first packet of a flow is always assigned to the first VOQ. To be precise, let $A_{i,\ell,j}^1(t)$ be the cumulative number of $A_{i,\ell}$-flow packets that are split into the j^{th} VOQ at the i^{th} input port of the first stage by time t. Then

$$A_{i,\ell,j}^1(t) = \lceil \frac{A_{i,\ell}(t) - j + 1}{N} \rceil, \quad j = 1, \dots, N, \tag{5}$$

and

$$A_{i,\ell}(t) = \sum_{j=1}^{N} A^1_{i,\ell,j}(t), \tag{6}$$

where $\lceil x \rceil$ (resp. $\lfloor x \rfloor$) is the ceiling (resp. floor) function of x.

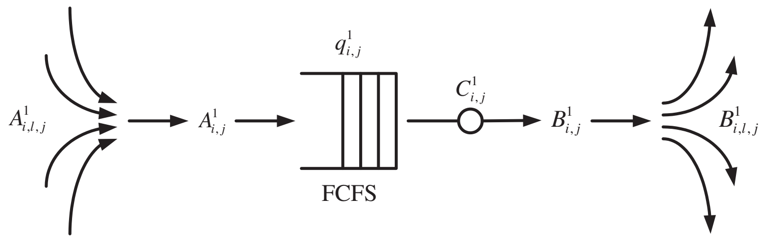

Fig. 5. The j^{th} VOQ at the i^{th} input port of the first stage

Now consider the j^{th} VOQ at the i^{th} input port of the first stage (see Figure 5). Let $A^1_{i,j}(t)$ be the cumulative number of arrivals by time t to this queue, $C^1_{i,j}(t)$ be its cumulative number of time slots assigned to this queue by time t, $q^1_{i,j}(t)$ be the number of packets queued at time t, and $B^1_{i,j}(t)$ be the cumulative number of departures by time t from this queue. Since such a queue can be viewed as a work conserving link with a time varying capacity under the FCFS policy, one has the following well-known representation (see e.g., [2])

$$q^1_{i,j}(t) = \max_{0 \le s \le t} [A^1_{i,j}(t) - A^1_{i,j}(s) - (C^1_{i,j}(t) - C^1_{i,j}(s))], \tag{7}$$

where

$$A^1_{i,j}(t) = \sum_{\ell=1}^{L_i} A^1_{i,\ell,j}(t). \tag{8}$$

Moreover,

$$B^1_{i,j}(t) = \min_{0 \le s \le t} [A^1_{i,j}(s) + C^1_{i,j}(t) - C^1_{i,j}(s)]. \tag{9}$$

In Lemma 3, we show that both the queue length and the delay at the load-balancing buffer can be bounded by finite constants. A similar result was previously shown in [9].

Lemma 3 ***(i)*** *The maximum queue length of the j^{th} VOQ at the i^{th} input port of the first stage is bounded above by L_i, i.e.,*

$$q^1_{i,j}(t) \le L_i, \quad \textit{for all}\, t. \tag{10}$$

(ii) *The maximum delay for a packet to depart the j^{th} VOQ at the i^{th} input port of the first stage is bounded above by $(N-1)L_i$.*

For the proof of Lemma 3, we need to introduce the following properies for the ceiling function and the floor function.

Proposition 4 ***(i)*** $\lceil a+b \rceil \le \lceil a \rceil + \lceil b \rceil \le \lceil a+b \rceil + 1.$
(ii) $\lfloor a+b \rfloor \ge \lfloor a \rfloor + \lfloor b \rfloor.$
(iii) $\lceil a \rceil \le \lfloor a \rfloor + 1.$
(iv) *For an integer a, and $N > 0$,*

$$\lceil \frac{a}{N} \rceil = \lfloor \frac{a+(N-1)}{N} \rfloor.$$

Proof. Since the proofs of (i), (ii) and (iii) are trivial, we only prove (iv). Without loss of generality, assume $\lceil \frac{a}{N} \rceil = P$ for some integer. Since a is an integer, we have

$$N(P-1)+1 \le a \le NP.$$

Adding $(N-1)$ yields

$$NP \le a+(N-1) \le NP+(N-1).$$

This then implies

$$\lfloor \frac{NP}{N} \rfloor \le \lfloor \frac{a+(N-1)}{N} \rfloor \le \lfloor \frac{N(P+1)-1}{N} \rfloor.$$

Thus,

$$\lfloor \frac{a+(N-1)}{N} \rfloor = P.$$

■

Proof. (Proof of Lemma 3) (i) Note from (8) and (5) that

$$A^1_{i,j}(t) - A^1_{i,j}(s) = \sum_{\ell=1}^{L_i} \Big(A^1_{i,\ell,j}(t) - A^1_{i,\ell,j}(s) \Big)$$
$$= \sum_{\ell=1}^{L_i} \Big(\lceil \frac{A_{i,\ell}(t)-j+1}{N} \rceil - \lceil \frac{A_{i,\ell}(s)-j+1}{N} \rceil \Big).$$

Applying the first inequality in Proposition 4 (i) yields

$$A^1_{i,j}(t) - A^1_{i,j}(s) \le \sum_{\ell=1}^{L_i} \lceil \frac{A_{i,\ell}(t) - A_{i,\ell}(s)}{N} \rceil. \tag{11}$$

Since the connection patterns at the first stage are periodic with period N for some one-cycle permutation matrix, we have

$$C^1_{i,j}(t) - C^1_{i,j}(s) \ge \lfloor \frac{t-s}{N} \rfloor. \tag{12}$$

Observe that $\sum_{\ell=1}^{L_i}(A_{i,\ell}(t) - A_{i,\ell}(s))$ is the number of packet arrivals at the i^{th} input port during the interval of length $t-s$. As there is at most one packet arrival at an input port per time slot, we have

$$\sum_{\ell=1}^{L_i}(A_{i,\ell}(t) - A_{i,\ell}(s)) \le t-s. \tag{13}$$

In conjunction with (12),

$$C^1_{i,j}(t) - C^1_{i,j}(s) \ge \lfloor \frac{\sum_{\ell=1}^{L_i}(A_{i,\ell}(t) - A_{i,\ell}(s))}{N} \rfloor \ge \sum_{\ell=1}^{L_i} \lfloor \frac{A_{i,\ell}(t) - A_{i,\ell}(s)}{N} \rfloor, \tag{14}$$

where we use Proposition 4 (ii) in the last inequality. From (7), (11), (14), and Proposition 4 (iii),

$$q^1_{i,j}(t) \le \max_{0 \le s \le t} \Big[\sum_{\ell=1}^{L_i} \Big(\lceil \frac{A_{i,\ell}(t) - A_{i,\ell}(s)}{N} \rceil - \lfloor \frac{A_{i,\ell}(t) - A_{i,\ell}(s)}{N} \rfloor \Big) \Big] \le L_i.$$

(ii) Since the scheduling policy at this queue is FCFS, it suffices to show that

$$B^1_{i,j}(t + (N-1)L_i) \ge A^1_{i,j}(t).$$

Note from (9) that

$$\begin{aligned} & B^1_{i,j}(t + (N-1)L_i) - A^1_{i,j}(t) \\ = & \min_{0 \le s \le t+(N-1)L_i} [A^1_{i,j}(s) - A^1_{i,j}(t) + C^1_{i,j}(t + (N-1)L_i) - C^1_{i,j}(s)] \\ = & \min\Big[\min_{0 \le s \le t} [C^1_{i,j}(t + (N-1)L_i) - C^1_{i,j}(s) - (A^1_{i,j}(t) - A^1_{i,j}(s))], \end{aligned}$$

$$\min_{t+1 \le s \le t+(N-1)L_i} \left[C^1_{i,j}(t+(N-1)L_i) - C^1_{i,j}(s) - (A^1_{i,j}(t) - A^1_{i,j}(s))\right]\Big].$$

All the terms in the second minimum are clearly nonnegative as both $A^1_{i,j}(t)$ and $C^1_{i,j}(t)$ are non-decreasing in t. On the other hand, for $0 \le s \le t$, we have from (12),(13), (11) and Proposition 4 (ii),(iv) that

$$\begin{aligned}
& C^1_{i,j}(t+(N-1)L_i) - C^1_{i,j}(s) - (A^1_{i,j}(t) - A^1_{i,j}(s)) \\
\ge & \lfloor \frac{t+(N-1)L_i - s}{N} \rfloor - (A^1_{i,j}(t) - A^1_{i,j}(s)) \\
\ge & \lfloor \frac{\sum_{\ell=1}^{L_i}(A_{i,\ell}(t) - A_{i,\ell}(s)) + (N-1)L_i}{N} \rfloor - \sum_{\ell=1}^{L_i} \lceil \frac{A_{i,\ell}(t) - A_{i,\ell}(s)}{N} \rceil \\
= & \lfloor \frac{\sum_{\ell=1}^{L_i}(A_{i,\ell}(t) - A_{i,\ell}(s) + (N-1))}{N} \rfloor - \sum_{\ell=1}^{L_i} \lceil \frac{A_{i,\ell}(t) - A_{i,\ell}(s)}{N} \rceil \\
\ge & \sum_{\ell=1}^{L_i} \Big(\lfloor \frac{A_{i,\ell}(t) - A_{i,\ell}(s) + (N-1)}{N} \rfloor - \lceil \frac{A_{i,\ell}(t) - A_{i,\ell}(s)}{N} \rceil \Big) = 0.
\end{aligned}$$

■

2.3 Analysis for the central buffer under FCFS

To ease the presentation, we let $d_{1,\max} = (N-1)L_{\max}$. Note from Lemma 3 that $d_{1,\max}$ is also the maximum delay at the first stage. As discussed in the introduction, a jitter control mechanism is added in the VOQ in front of the second stage. As the delay through the load-balancing buffer is bounded above by $d_{1,\max}$, packets with delay less than $d_{1,\max}$ are delayed to $d_{1,\max}$. By so doing, every packet has the same delay before entering the buffer at an input port of the second stage.

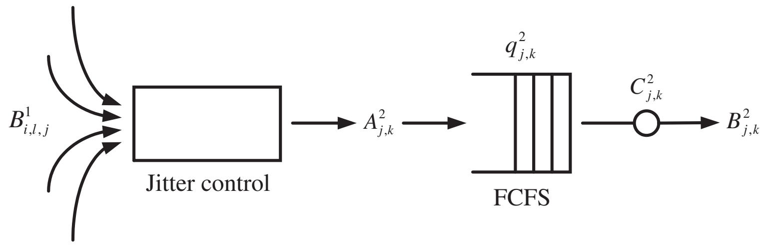

Fig. 6. The k^{th} VOQ at the j^{th} input port of the second stage under FCFS

Now consider the k^{th} VOQ at the j^{th} input port of the second stage (see Figure 6). Let $A^2_{j,k}(t)$ be the cumulative number of arrivals by time t to this queue, $C^2_{j,k}(t)$ be its cumulative number of time slots assigned to this queue by time t, $q^2_{j,k}(t)$ be the number of packets queued at time t, and $B^2_{j,k}(t)$ be the cumulative number of departures by time t from this queue. Since every packet has the same delay $d_{1,\max}$ through the first stage and fan-out spiltting is done at this stage, we have

$$A^2_{j,k}(t) = \sum_{(i,\ell)\in S^*(k)} A^1_{i,\ell,j}(t - d_{1,\max}) \tag{15}$$

(Here we use the convention that $A^1_{i,\ell,j}(\tau) = 0$ for $\tau < 0$). Also, as this queue is a work conserving link with a time varying capacity under the FCFS policy, one has

$$q^2_{j,k}(t) = \max_{0\le s\le t}[A^2_{j,k}(t) - A^2_{j,k}(s) - (C^2_{j,k}(t) - C^2_{j,k}(s))], \tag{16}$$

and

$$B^2_{j,k}(t) = \min_{0\le s\le t}[A^2_{j,k}(s) + C^2_{j,k}(t) - C^2_{j,k}(s)]. \tag{17}$$

Lemma 5 *For the FCFS scheme,*

(i) $q^2_{j,k}(t) \le \lceil \frac{q^o_k(t-d_{1,\max})}{N} \rceil + M_k$*, for all* t*, and*
(ii) $B^2_{j,k}(t + q^o_k(t - d_{1,\max}) + (N-1)M_k) \ge A^2_{j,k}(t)$*, for all* t*.*

Lemma 5 (i) provides an upper bound for the queue length in terms of the queue length of the corresponding output-buffered switch. As the scheduling policy is FCFS, Lemma 5 (ii) implies that a packet that arrives at the queue at time t will depart not later than $t + q^o_k(t - d_{1,\max}) + (N-1)M_{\max}$.

Proof. (i) Note from (15),(5) and the inequality in Proposition 4 (i) that

$$A^2_{j,k}(t) - A^2_{j,k}(s) = \sum_{(i,\ell)\in S^*(k)} (A^1_{i,\ell,j}(t - d_{1,\max}) - A^1_{i,\ell,j}(s - d_{1,\max}))$$

$$\le \sum_{(i,\ell)\in S^*(k)} \lceil \frac{A_{i,\ell}(t - d_{1,\max}) - A_{i,\ell}(s - d_{1,\max})}{N} \rceil \tag{18}$$

$$\le \lceil \frac{\sum_{(i,\ell)\in S^*(k)}(A_{i,\ell}(t - d_{1,\max}) - A_{i,\ell}(s - d_{1,\max}))}{N} \rceil + M_k - 1. \tag{19}$$

Observe from (1) and (2) that

$$\sum_{(i,\ell)\in S^*(k)} (A_{i,\ell}(t - d_{1,\max}) - A_{i,\ell}(s - d_{1,\max}))$$

$$
\begin{aligned}
&= A_k^o(t-d_{1,\max}) - A_k^o(s-d_{1,\max}) \\
&\le q_k^o(t-d_{1,\max}) + (t-s).
\end{aligned} \tag{20}
$$

Thus,

$$
A_{j,k}^2(t) - A_{j,k}^2(s) \le \lceil \frac{q_k^o(t-d_{1,\max}) + (t-s)}{N} \rceil + M_k - 1. \tag{21}
$$

Since the connection patterns at the second stage are also periodic with period N for some one-cycle permutation matrix,

$$
C_{j,k}^2(t) - C_{j,k}^2(s) \ge \lfloor \frac{t-s}{N} \rfloor. \tag{22}
$$

From (21), (22), (16) and Proposition 4 (iii), it then follows that

$$
q_{j,k}^2(t) \le \lceil \frac{q_k^o(t-d_{1,\max})}{N} \rceil + M_k.
$$

for all t.
(ii) Let $d = q_k^o(t-d_{1,\max}) + (N-1)M_k$. Note from (17) that

$$
\begin{aligned}
&B_{j,k}^2(t) - A_{j,k}^2(t) \\
&= \min_{0\le s\le t+d} [A_{j,k}^2(s) - A_{j,k}^2(t) + C_{i,j}^2(t+d) - C_{i,j}^2(s)] \\
&= \min\Big[\min_{0\le s\le t} [C_{j,k}^2(t+d) - C_{j,k}^2(s) - (A_{j,k}^2(t) - A_{j,k}^2(s))], \\
&\quad \min_{t+1\le s\le t+d} [C_{j,k}^2(t+d) - C_{j,k}^2(s) - (A_{j,k}^2(t) - A_{j,k}^2(s))]\Big]
\end{aligned}
$$

Clearly, all the terms in the second minimum are nonnegative as both $A_{j,k}^2(t)$ and $C_{j,k}^2(t)$ are nondecreasing in t. On the other hand, for $0 \le s \le t$, we have from (18) that

$$
A_{j,k}^2(t) - A_{j,k}^2(s) \le \sum_{(i,\ell)\in S^*(k)} \lceil \frac{A_{i,\ell}(t-d_{1,\max}) - A_{i,\ell}(s-d_{1,\max})}{N} \rceil. \tag{23}
$$

Also, it follows from (22), (20), and proposition 4 (ii) that

$$
\begin{aligned}
&C_{j,k}^2(t+d) - C_{j,k}^2(s) \ge \lfloor \frac{t+d-s}{N} \rfloor \\
&= \lfloor \frac{q_k^o(t-d_{1,\max}) + (t-s) + (N-1)M_k}{N} \rfloor
\end{aligned}
$$

$$
\begin{aligned}
&\geq \lfloor \frac{\sum_{(i,\ell)\in S^*(k)}(A_{i,\ell}(t-d_{1,\max})-A_{i,\ell}(s-d_{1,\max}))+(N-1)M_k}{N} \rfloor \\
&= \lfloor \frac{\sum_{(i,\ell)\in S^*(k)}(A_{i,\ell}(t-d_{1,\max})-A_{i,\ell}(s-d_{1,\max})+(N-1))}{N} \rfloor \\
&\geq \sum_{(i,\ell)\in S^*(k)} \lfloor \frac{A^2_{j,k}(t-d_{1,\max})-A^2_{j,k}(s-d_{1,\max})+(N-1)}{N} \rfloor . \qquad (24)
\end{aligned}
$$

As in the proof of Lemma 3, we then have from (23), (24) and Proposition 4 (iv) that

$$C^2_{j,k}(t+d)-C^2_{j,k}(s)-(A^2_{j,k}(t)-A^2_{j,k}(s)) \geq 0.$$

for $0 \leq s \leq t$. ■

2.4 Analysis for the resequencing-and-output buffer

In this section, we analyze the resequencing-and-output buffer. The resequencing-and-output buffer conceptually consists of two virtual buffers (see Figure 7): (i) the resequencing buffer and (ii) the output buffer. The objective of the resequencing buffer is to reorder the packets so that packets of the same flow depart in the same order as they arrive. After resequencing, packets are stored in the output buffer waiting for transmission from the output link. Let $A^3_k(t)$ be the cumulative arrivals by time t to the k^{th} resequencing buffer, and $B^3_k(t)$ be its cumulative departures. Note that $B^3_k(t)$ is also the cumulative arrivals by time t to the k^{th} output buffer. Let $B^4_k(t)$ be the cumulative departures by time t from the k^{th} output buffer. Clearly, from the input-output relation of a work-conserving link, we have

$$B^4_k(t) = \min_{0\leq s\leq t}[B^3_k(s)+(t-s)]. \qquad (25)$$

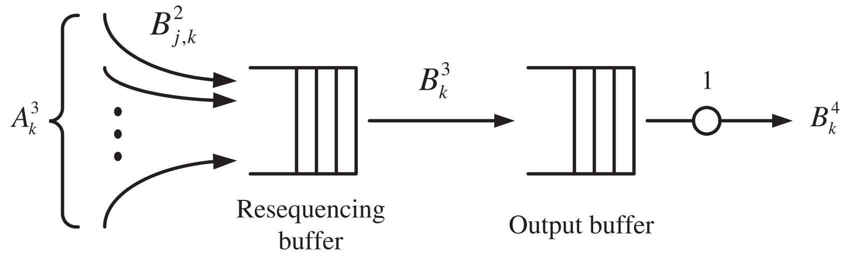

Fig. 7. The resequencing-and-output buffer

Lemma 6 *The following results hold for the FCFS scheme.*

(i) $A^3_k(t) \leq B^o_k(t - d_{1,\max})$.
(ii) $B^3_k(t + d_{1,\max} + NM_k) \geq B^o_k(t)$.
(iii) $B^4_k(t) \geq B^o_k(t - d_{1,\max} - NM_k)$.
(iv) *The number of packets queued at the* k^{th} *resequencing-and-output buffer is bounded above by* NM_k*, i.e.,*

$$A^3_k(t) - B^4_k(t) \leq NM_k, \quad \textit{for all}\, t. \tag{26}$$

Proof. (i) Note that

$$A^3_k(t) = \sum_{j=1}^{N} B^2_{j,k}(t).$$

From (17), it follows that

$$A^3_k(t) = \sum_{j=1}^{N} \min_{0 \leq s \leq t} [A^2_{j,k}(s) + C^2_{j,k}(t) - C^2_{j,k}(s)]$$
$$\leq \min_{0 \leq s \leq t} [\sum_{j=1}^{N} A^2_{j,k}(s) + C^2_{j,k}(t) - C^2_{j,k}(s)].$$

As the connection patterns at the second stage are periodic with period N for some one-cycle permutation matrix,

$$\sum_{j=1}^{N} C^2_{j,k}(t) = t.$$

Thus,

$$A^3_k(t) \leq \min_{0 \leq s \leq t} [\sum_{j=1}^{N} A^2_{j,k}(s) + t - s]. \tag{27}$$

Note from (15), (6), and (2) that for all s

$$\sum_{j=1}^{N} A^2_{j,k}(s) = \sum_{j=1}^{N} \sum_{(i,\ell) \in S^*(k)} A^1_{i,\ell,j}(s - d_{1,\max})$$
$$= \sum_{(i,\ell) \in S^*(k)} A_{i,\ell}(s - d_{1,\max})$$
$$= A^o_k(s - d_{1,\max}).$$

Using this and (3) in (27) yields

$$\begin{aligned}
A_k^3(t) &\le \min_{0\le s\le t}[A_k^o(s-d_{1,\max})+t-s] \\
&= \min_{d_{1,\max}\le s\le t}[A_k^o(s-d_{1,\max})+t-s] \\
&= \min_{0\le \tau\le t-d_{1,\max}}[A_k^o(\tau)+t-d_{1,\max}-\tau] \\
&= B_k^o(t-d_{1,\max})
\end{aligned}$$

(ii) Consider a packet that is destined for the k^{th} output port. Without loss of generality, suppose that this packet arrives at time t and it is routed through the k^{th} VOQ at j^{th} input port of the second stage. From Lemma 3(ii), the packet leaves the first stage not later than $t+d_{1,\max}$. After the jitter control, it arrives at the j^{th} input port of the second stage exactly at $t+d_{1,\max}$. Since FCFS is used in the k^{th} VOQ at the j^{th} input port of the second stage, we have from Lemma 5(ii) that the packet leaves the second stage not later than $t+q_k^o(t)+(N-1)M_k+d_{1,\max}$. As the bound is independent of j, we conclude that every packet that arrives at the first stage at time t will depart from the second stage not later than $t+q_k^o(t)+(N-1)M_k+d_{1,\max}$. Note from (1) that $t+q_k^o(t)$ is non-decreasing in t. Thus, any other packets that are destined for the k^{th} output port and arrives before t leave the second stage not later than $t+q_k^o(t)+(N-1)M_k+d_{1,\max}$. This implies the packet (that arrives at the first stage at time t and destined for the k^{th} output port) leaves the resequencing buffer not later than $t+q_k^o(t)+(N-1)M_k+d_{1,\max}$.

On the other hand, since the packet that arrives at time t departs from the corresponding output-buffered switch between $t+q_k^o(t)-M_k+1$ and $t+q_k^o(t)$, the departure time for a packet to leave the resequencing buffer is not later than the sum of that from the corresponding output-buffer switch and $d_{1,\max}+NM_k$. This shows that for all t

$$B_k^3(t+d_{1,\max}+NM_k)\ge B_k^o(t).$$

(iii) From (25) and (ii) of this lemma, it follows that

$$B_k^4(t)\ge \min_{0\le s\le t}[B_k^o(s-d_{1,\max}-NM_k)+t-s].$$

Since $B_k^o(t)-B_k^o(s)\le t-s$ in (4), the above minimum occurs at $s=t$. Thus,

$$B_k^4(t)\ge B_k^o(t-d_{1,\max}-NM_k).$$

(iv) From (i) and (iii) of this lemma and (4), it follows that

$$A_k^3(t) - B_k^4(t) \leq B_k^o(t - d_{1,\max}) - B_k^o(t - d_{1,\max} - NM_k)$$
$$\leq NM_k.$$

■

Proof. (Proof of Theorem 1) (i) This is a direct consequence of Lemma 6(iii).

(ii) Since there are N VOQ at each input port, the result then follows from Lemma 3(i).

(iii) This is also shown in Lemma 3(ii).

(iv) It is shown in Lemma 6(iv). ■

3 The EDF scheme

In this section, we prove Theorem 2. Consider the same traffic model as in Section 2. Also, let $B_{i,\ell}^o(t)$ be the cumulative departures of the $A_{i,\ell}$-flow by time t from the corresponding FCFS output-buffered switch (see Figure 4). Since the EDF scheme uses packet departure times as deadlines, $B_{i,\ell}^o(t)$ is also the number of packets from the $A_{i,\ell}$-flow that have deadlines not greater than t. Note that

$$B_k^o(t) = \sum_{(i,\ell)\in S^*(k)} B_{i,\ell}^o(t). \tag{28}$$

We first establish the following inequalities that will be used in the proof of Theorem 2.

Proposition 7 *For the FCFS output-buffered switch,*

$$\sum_{(i,\ell)\in S(k)} B_{i,\ell}^o(t) \leq \min_{0\leq s\leq t} [\sum_{(i,\ell)\in S(k)} A_{i,\ell}(s) + t - s], \tag{29}$$

for all t and for all $S(k)$ that is a subset of $S^*(k)$.

Proof. If one assigns priority to the set of traffic from $S(k)$ in the output-buffered switch, then the right hand side of (29) is the cumulative departures by time t for this set of traffic (cf. (3)). This should be larger than or equal to the left hand of (29), which is the cumulative departures by time t for this set of traffic under FCFS. ■

We also note that results in Section 2.2 can be directly applied as the operations for the load-balancing buffer in front of the first stage in the EDF scheme is exactly the same as that in Section 2.2.

3.1 Analysis for the central buffer under EDF

Now consider the k^{th} VOQ at the j^{th} input port of the second stage (see Figure 8). As in Section 2.3, let $A^2_{j,k}(t)$ be the cumulative number of arrivals by time t to this queue and $C^2_{j,k}(t)$ be its cumulative number of time slots assigned to this queue by time t.

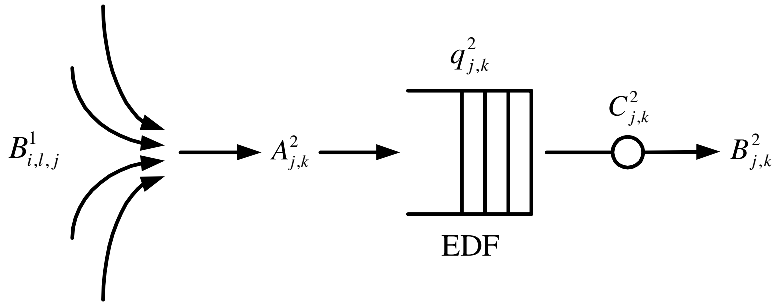

Fig. 8. The k^{th} VOQ at the j^{th} input port of the second stage under EDF

Also, let $B^1_{i,\ell,j}$ be the departure process of $A^1_{i,\ell,j}$ from the j^{th} VOQ at the i^{th} input port of the first stage. Since there is no jitter control in the EDF scheme, the departure process from the the first stage is simply the arrival process at the second stage. Thus,

$$A^2_{j,k}(t) = \sum_{(i,\ell)\in S^*(k)} B^1_{i,\ell,j}(t). \tag{30}$$

Lemma 8 *Every packet leaves the k^{th} VOQ at the j^{th} input port of the second stage not later than the sum of its deadline and $(N-1)(L_{\max}+M_k)$.*

Note that if we add the deadline of every packet by the same constant $(N-1)(L_{\max}+M_k)$, then Lemma 8 is equivalent to that every packet leaves the k^{th} VOQ at the j^{th} input port of the second stage not later than its deadline.

Proof. Let $d_1 = (N-1)L_{\max}$ and $d_2 = (N-1)M_k$. Let $D_{i,\ell,j}(t)$ be the number of packets of the $A^1_{i,\ell,j}$-flow that have deadlines not greater than t. From [2], Theorem 5.6.1, it suffices to show that

$$
\begin{aligned}
&\sum_{(i,\ell)\in S(k)} D_{i,\ell,j}(t-d_1-d_2)\\
&\leq \min_{0\leq s\leq t}[\sum_{(i,\ell)\in S(k)} B^1_{i,\ell,j}(s)+C^2_{j,k}(t)-C^2_{j,k}(s)],
\end{aligned}
\tag{31}
$$

for all t and for all $S(k)$ that is a subset of $S^*(k)$.

As the delay through the j^{th} VOQ at an input port of the first stage is bounded above by $d_1=(N-1)L_{\max}$ in Lemma 3(ii), we then have

$$
\begin{aligned}
&\min_{0\leq s\leq t}[\sum_{(i,\ell)\in S(k)} B^1_{i,\ell,j}(s)+C^2_{j,k}(t)-C^2_{j,k}(s)]\\
&\geq \min_{0\leq s\leq t}[\sum_{(i,\ell)\in S(k)} A^1_{i,\ell,j}(s-d_1)+C^2_{j,k}(t)-C^2_{j,k}(s)]\\
&= \min_{0\leq \tau\leq t-d_1}[\sum_{(i,\ell)\in S(k)} A^1_{i,\ell,j}(\tau)+C^2_{j,k}(t)-C^2_{j,k}(\tau+d_1)].
\end{aligned}
$$

Using (5) and (22), one has

$$
\begin{aligned}
&\min_{0\leq s\leq t}[\sum_{(i,\ell)\in S(k)} B^1_{i,\ell,j}(s)+C^2_{j,k}(t)-C^2_{j,k}(s)]\\
&\geq \min_{0\leq \tau\leq t-d_1}\Big[\sum_{(i,\ell)\in S(k)} \lceil\frac{A_{i,\ell}(\tau)-j+1}{N}\rceil+\lfloor\frac{t-d_1-\tau}{N}\rfloor\Big].
\end{aligned}
\tag{32}
$$

On the other hand, since we use the departure times from the corresponding output-buffered switch as deadlines and packets are split in the round robin fashion, we have (cf. (5))

$$
D_{i,\ell,j}(t)=\lceil\frac{B^o_{i,\ell}(t)-j+1}{N}\rceil.
\tag{33}
$$

In conjunction with (32), the inequalities in (31) hold if we can show that for all $0\leq\tau\leq t-d_1$,

$$
\begin{aligned}
&\sum_{(i,\ell)\in S(k)}\lceil\frac{B^o_{i,\ell}(t-d_1-d_2)-j+1}{N}\rceil\\
&\leq \sum_{(i,\ell)\in S(k)}\lceil\frac{A_{i,\ell}(\tau)-j+1}{N}\rceil+\lfloor\frac{t-d_1-\tau}{N}\rfloor.
\end{aligned}
\tag{34}
$$

Observe from Proposition 7 that

$$\sum_{(i,\ell)\in S(k)} B^o_{i,\ell}(t-d_1-d_2) \le \sum_{(i,\ell)\in S(k)} A_{i,\ell}(\tau) + t - d_1 - d_2 - \tau, \tag{35}$$

for all τ, t and all subsets $S(k)$. In particular, choosing $\tau = t - d_1 - d_2$ and a subset that contains only one flow yields

$$A_{i,\ell}(t-d_1-t_2) \ge B^o_{i,\ell}(t-d_1-d_2),$$

for all $(i,\ell) \in S^*(k)$. Thus, for $t-d_1-d_2+1 \le \tau \le t-d_1$, one has

$$A_{i,\ell}(\tau) \ge A_{i,\ell}(t-d_1-t_2) \ge B^o_{i,\ell}(t-d_1-d_2),$$

and the inequality in (34) is satisfied trivially.

For $0 \le \tau \le t-d_1-d_2$, we have from the inequality (35) that

$$t-d_1-\tau \ge \sum_{(i,\ell)\in S(k)} \Big(B^o_{i,\ell}(t-d_1-d_2) - A_{i,\ell}(\tau)\Big) + d_2.$$

Recall that $d_2 = (N-1)M_k$. Also, we have $|S(k)| \le M_k$ as $S(k)$ is a subset of $S^*(k)$. Thus,

$$\begin{aligned}
&\lfloor\frac{t-d_1-\tau}{N}\rfloor \ge \lfloor\frac{\sum_{(i,\ell)\in S(k)}\Big(B^o_{i,\ell}(t-d_1-d_2) - A_{i,\ell}(\tau)\Big) + (N-1)M_k}{N}\rfloor \\
&\ge \lfloor\frac{\sum_{(i,\ell)\in S(k)}\Big(B^o_{i,\ell}(t-d_1-d_2) - A_{i,\ell}(\tau)\Big) + (N-1)|S(k)|}{N}\rfloor \\
&= \lfloor\frac{\sum_{(i,\ell)\in S(k)}\Big(B^o_{i,\ell}(t-d_1-d_2) - A_{i,\ell}(\tau) + (N-1)\Big)}{N}\rfloor.
\end{aligned}$$

Applying Proposition 4 (ii) yields

$$\lfloor\frac{t-d_1-\tau}{N}\rfloor \ge \sum_{(i,\ell)\in S(k)} \lfloor\frac{B^o_{i,\ell}(t-d_1-d_2) - A_{i,\ell}(\tau) + (N-1)}{N}\rfloor. \tag{36}$$

On the other hand, we also have from Proposition 4 (i) that

$$\sum_{(i,\ell)\in S(k)} \lceil\frac{B^o_{i,\ell}(t-d_1-d_2) - j + 1}{N}\rceil - \sum_{(i,\ell)\in S(k)} \lceil\frac{A_{i,\ell}(\tau) - j + 1}{N}\rceil$$

$$\leq \sum_{(i,\ell)\in S(k)} \lceil \frac{B^o_{i,\ell}(t-d_1-d_2)-A_{i,\ell}(\tau)}{N} \rceil. \tag{37}$$

The inequalities in (34) then follows from (36), (37) and Proposition 4 (iv). ■

3.2 Analysis for the resequencing-and-output buffer under EDF

In this section, we analyze the resequencing-and-output buffer under the EDF scheme. As in Section 2.4, let $A^3_k(t)$ be the cumulative arrivals by time t to the k^{th} resequencing buffer, $B^3_k(t)$ be its cumulative departures, and $B^4_k(t)$ be the cumulative departures by time t from the k^{th} output buffer (see Figure 7).

Lemma 9 *The following results hold for the EDF scheme.*

(i) $A^3_k(t) \leq B^o_k(t)$.
(ii) $B^3_k(t+(N-1)(L_{\max}+M_k)) \geq B^o_k(t)$.
(iii) $B^4_k(t) \geq B^o_k(t-(N-1)(L_{\max}+M_k))$.
(iv) *The number of packets queued at the k^{th} resequencing-and-output buffer is bounded above by $(N-1)(L_{\max}+M_k)$, i.e.,*

$$A^3_k(t) - B^4_k(t) \leq (N-1)(L_{\max}+M_k), \quad \textit{for all}\, t. \tag{38}$$

Proof. (i) Note from (30), $B^1_{i,\ell,j}(s) \leq A^1_{i,\ell,j}(s)$, (6) and (2) that for all s

$$\sum_{j=1}^{N} A^2_{j,k}(s) = \sum_{j=1}^{N} \sum_{(i,\ell)\in S^*(k)} B^1_{i,\ell,j}(s) \leq \sum_{j=1}^{N} \sum_{(i,\ell)\in S^*(k)} A^1_{i,\ell,j}(s)$$
$$= \sum_{(i,\ell)\in S^*(k)} A_{i,\ell}(s) = A^o_k(s).$$

As the EDF scheme is work conserving, (27) in the proof of Lemma 6 is still applicable. It then follows from (3) that

$$A^3_k(t) \leq \min_{0\leq s\leq t}[A^o_k(s)+t-s] = B^o_k(t).$$

(ii) Since we use the departure times from the corresponding output-buffered switch as deadlines, the result then follows from Lemma 8.

(iii) This follows the same argument in Lemma 6(iii).

(iv) From (i) and (iii) of this lemma and (4), it follows that

$$
\begin{aligned}
&A_k^3(t) - B_k^4(t) \le B_k^o(t) - B_k^o(t - (N-1)(L_{\max} + M_k)) \\
&\le (N-1)(L_{\max} + M_k).
\end{aligned}
$$

■

Proof. (Proof of Theorem 2) (i) This is a direct consequence of Lemma 9(iii).

(ii) It is shown in Lemma 9(iv). ■

4 Conclusions

In this paper, we showed that resequencing can be done efficiently by distributing packets according to their flows in the load balanced Birkhoff-von Neumann switches. This is done by adding a load-balancing buffer in front of the first stage and a resequencing-and-output buffer after the second stage. We considered two scheduling policies: the FCFS policy and the EDF policy. For both schemes, we showed that the end-to-end delay is bounded by the sum of the delay from the corresponding FCFS output-buffered switch and a constant that only depends on the number of flows and the size of the switch. The FCFS policy requires jitter control in front of the second stage so that packets entering the second stage have the same delay. Its scheduling complexity is simpler than that of the EDF scheme. The EDF policy uses the departure times of the corresponding output-buffered switch as deadlines and schedules packets according to their deadlines. As such, there is no need for jitter control. Since computing departure times of the corresponding output-buffered switch requires the global information of all the inputs, we also suggested using the arrival times as deadlines for the EDF scheme. This reduces the scheduling complexity at the cost of slightly increasing the delay bound and the size of the resequencing-and-output buffer.

There are several problems that require further study.

(i) We only looked at the worst case of the end-to-end delay. It is of the same importance to study the average delay. In particular, the jitter control mechanism in the FCFS scheme that delays every packet to its maximum (worst case) delay is expected to perform poorly in the light load (the main reason that we have jitter control in the FCFS scheme is for the sake of the proof for the worst case bound). Our preliminary simulation study shows that the average delay with jitter control is mainly dominated by the delay in the jitter control even when the load is moderate.

(ii) Though we only compared our schemes with the FCFS output-buffered switches, we expect that our approaches could also be used for other work conserving scheduling policies, such as the Service Curve Earliest Deadline first (SCED) policy in [7] and the Generalized Processor Sharing (GPS) policy in [11]. For both SCED and GPS, it is well-known that Quality of Service (QoS) guarantees can be achieved in output-buffered switches. If there is a bound for the difference between our multi-stage switch and the corresponding output-buffered switch under these two policies, then QoS guarantees can also be achieved in our multi-stage switch. Research along this line will be reported separately.

(iii) In this paper, packets are assumed to be of the same size. To allow variable length packets, one needs to introduce packetizers (see e.g., [2]) for segmentation and reassembly. The bounds in [2] for packetizers might be used to extend the results in this paper to the case with variable length packets.

(iv) In the recent paper by Keslassy and McKeown [10], a "full frame first" algorithm is proposed so that packets in the two-stage switches can be departed in sequence. However, this is at the cost of communications between VOQs. It would be interesting to see if there are other scheduling algorithms that achieve the same objective with lower communication cost.

References

[1] G. Birkhoff, "Tres observaciones sobre el algebra lineal," *Univ. Nac. Tucumán Rev. Ser. A*, Vol. 5, pp. 147-151, 1946.

[2] C.S. Chang. *Performance Guarantees in Communication Networks.* London: Springer-Verlag, 2000.

[3] C.S. Chang, W.J. Chen and H.Y. Huang, "On service guarantees for input buffered crossbar switches: a capacity decomposition approach by Birkhoff and von Neumann," *IEEE IWQoS'99*, pp. 79-86, London, U.K., 1999 (U.S. patent pending).

[4] C.S. Chang, W.J. Chen and H.Y. Huang, "Birkhoff-von Neumann input buffered crossbar switches," *IEEE INFOCOM2000*, pp. 1614-1623, Tel Aviv, Israel, 2000.

[5] C.S. Chang, D.S. Lee and Y.S. Jou, "Load Balanced Birkhoff-von Neumann Switches, Part I: One-stage Buffering," to appear in the special issue of *Computer Communications* on "Current Issues in Terabit Switching," 2001.

[6] S.-T. Chuang, A. Goel, N. McKeown and B. Prabhkar, "Matching output queueing with a combined input output queued switch," *IEEE INFOCOM'99*, pp. 1169-1178, New York, 1999.

[7] R. L. Cruz, "Quality of service guarantees in virtual circuit switched networks," *IEEE Journal of Selected Areas in Communication, special issue on "Advances in the Fundamentals of Networking"*, August, 1995.

[8] J.Y. Hui and T. Renner, "Queueing strategies for multicast packet switching," *IEEE GLOBECOM'90*, Vol. 3, pp. 1431-1437, 1990.

[9] S. Iyer and N. McKeown, "Making parallel packet switch practical," *Proc. IEEE INFOCOM 2001*, Anchorage, Alaska, U.S.A.

[10] I. Keslassy and N. McKeown, "Maintaining packet order in two-stage switches," *preprint*, 2001.

[11] A.K. Parekh and R.G. Gallager, "A generalized processor sharing approach to flow control in integrated service networks: the single-node case," *IEEE/ACM Transactions on Networking*, Vol. 1, pp. 344-357, 1993.

[12] I. Stoica and H. Zhang, "Exact emulation of an output queueing switch by a combined input output queueing switch," *IEEE IWQoS'98*, pp. 218-224, Napa, California, 1998.

[13] J. von Neumann, "A certain zero-sum two-person game equivalent to the optimal assignment problem, " *Contributions to the Theory of Games,* Vol. 2, pp. 5-12, Princeton University Press, Princeton, New Jersey, 1953.