Quasi-Output-Buffered Switches

Cheng-Shang Chang, *Fellow, IEEE*, Jay Cheng, *Senior Member, IEEE*,
Duan-Shin Lee, *Senior Member, IEEE*, and Chi-Feung Wu

***Abstract*—It is well known that output-buffered switches have better performance than other switch architectures. However, output-buffered switches also suffer from the notorious scalability problem, and direct constructions of large output-buffered switches are difficult. In this paper, we study the problem of constructing *scalable* switches that have comparable performance (in the sense of 100% throughput and first-in first-out (FIFO) delivery of packets from the same flow) to output-buffered switches. For this, we propose a new concept, called *quasi-output-buffered switch*. Like an output-buffered switch, a quasi-output-buffered switch is a *deterministic* switch that *achieves 100% throughput* and *delivers packets from the same flow in the FIFO order*. Using the three-stage Clos network, we show that one can *recursively* construct a larger quasi-output-buffered switch with a set of smaller quasi-output-buffered switches. By recursively expanding the three-stage Clos network, we obtain a quasi-output-buffered switch with only 2×2 switches. Such a switch is called a *packet-pair switch* in this paper as it always transmits packets in pairs. By computer simulations, we show that packet-pair switches have better delay performance than most load-balanced switches with comparable construction complexity.**

***Index Terms*—Delay performance, load-balanced switches, output-buffered switches, packet-pair switches, quasi-output-buffered switches.**

I. Introduction

It is well known that output-buffered switches achieve 100% throughput and have the best delay performance among all switch architectures. However, this is at the cost of N times speedup for an $N \times N$ output-buffered switch. The required speedup makes it difficult to construct a large output-buffered switch. There are several studies in the literature that achieve exact emulation of an output-buffered switch, such as the crosspoint-buffered switch [1], the parallel-buffered switch [2], and the combined input/output-buffered switch [3], [4]. However, all of these switches either have non-scalable hardware complexity or have computation and communication overheads.

One of the key problems in high speed switching is whether one can construct *scalable* switches with comparable performance to output-buffered switches. Recent advances in load-balanced switches (see e.g., [5]–[9]) have shed some light on that problem. A typical load-balanced switch consists of two stages: the first stage is for load balancing that converts incoming traffic into uniform traffic, and the second stage is for switching of the uniform traffic. Moreover, the connection patterns for the crossbar switches in the two stages of a load-balanced switch are *deterministic* and *periodic*. It is shown that various load-balanced switches have comparable performance to output-buffered switches. As such, they can achieve 100% throughput with $O(1)$ computation and communication overheads.

One of the main contributions of this paper is to identify the key ingredients in load-balanced switches that enable us to construct large switches with comparable performance (in the sense of 100% throughput and first-in first-out (FIFO) delivery of packets from the same flow) to output-buffered switches. For this, we propose a new concept, called *quasi-output-buffered switch*. Like an output-buffered switch, a quasi-output-buffered switch is a *deterministic* switch that *achieves 100% throughput* and *delivers packets from the same flow in the FIFO order*. Using the three-stage Clos network [10], we show that one can *recursively* construct a larger quasi-output-buffered switch with a set of smaller quasi-output-buffered switches. To the best of our knowledge, such a result on quasi-output-buffered switches seems to be the first result that allows *recursive constructions* of switches with comparable performance (in the sense of 100% throughput and FIFO delivery of packets from the same flow) to output-buffered switches. Analogous to the construction of a Benes network [11], we recursively expand the three-stage Clos network to obtain a quasi-output-buffered switch with only 2×2 switches. Such a switch is called a *packet-pair switch* in this paper as it always transmits packets in pairs. The packet-pair switches have several nice features: 100% throughput, FIFO delivery of packets from the same flow, deterministic connection patterns of all 2×2 switches, self-routing of packets, and no need for computation and communication. By computer simulations, we also show that packet-pair switches have better delay performance than most load-balanced switches with comparable construction complexity.

The key theory behind our constructions of quasi-output-buffered switches is a refined *calculus* based on a traffic characterization in [12]. Such a traffic characterization allows us to describe a flow of packets by a single "rate." We show that the aggregated flow has a rate equal to the sum of the rates of individual flows. Round-robin splitting of a flow yields several subflows with smaller rates. Moreover, a departing flow has the same rate as that of the arriving flow provided that the system is "stable." Unlike the theory of effective bandwidth

This research was supported in part by the National Science Council, Taiwan, R.O.C., under Contract NSC-95-2221-E-007-039, NSC-96-2221-E-007-076, NSC-97-2221-E-007-105-MY3, and the Program for Promoting Academic Excellence of Universities NSC 94-2752-E-007-002-PAE. This paper was presented in part at the IEEE Annual Conference on Computer Communications (INFOCOM'08), Phoenix, AZ, USA, April 13–18, 2008.

Cheng-Shang Chang and Chi-Feung Wu are with the Institute of Communications Engineering, National Tsing Hua University, Hsinchu 30013, Taiwan, R.O.C. (e-mail: cschang@ee.nthu.edu.tw; cfwu@gibbs.ee.nthu.edu.tw).

Jay Cheng is with the Department of Electrical Engineering and the Institute of Communications Engineering, National Tsing Hua University, Hsinchu 30013, Taiwan, R.O.C. (e-mail: jcheng@ee.nthu.edu.tw).

Duan-Shin Lee is with the Department of Computer Science and the Institute of Communications Engineering, National Tsing Hua University, Hsinchu 30013, Taiwan, R.O.C. (e-mail: lds@cs.nthu.edu.tw).

(see e.g., [13] and the references therein), the refined calculus does not need the *independence* assumption on the flows.

The paper is organized as follows. In Section II, we introduce the traffic characterization and its associated calculus. We also define the concept of a quasi-output-buffered switch in Section II. Then we propose a three-stage construction of a quasi-output-buffered switch in Section III and introduce the packet-pair switches in Section IV. Section V concludes this paper.

II. Quasi-Output-Buffered Switches

In this paper, we only consider the discrete-time setting and we make the following assumptions: (i) Time is slotted and synchronized in every link. (ii) Packets are of the same size and each packet can be transmitted within a time slot. A *flow* is commonly known as a sequence of packets that have the same source-destination pair in a switch (or a network of switches).

In Section II-A, we first give the traffic characterization for the flows of packets considered in this paper, and show that such flows possess three useful properties: the superposition property, the splitting property, and the departure property. In Section II-B, we review the definition of an output-buffered switch and show that output-buffered switches have the universal stability property under a no overbooking condition for the input traffic flows. By extracting and preserving some key properties in output-buffered switches, we then formally define a quasi-output-buffered switch in Section II-C, and show in Section II-D that the key properties of quasi-output-buffered switches can be preserved through a feedforward network if the total "mean" arrival rate at every output link of every quasi-output-buffered switch in the feedforward network does not exceed its capacity. Such a result will be useful in the three-stage construction of a quasi-output-buffered switch in Section III.

A. Traffic Characterization

In most switching papers, traffic characterizations for flows in a switch (or a network of switches) are usually assumed to follow certain traffic models, e.g., Bernoulli arrival processes and Markov processes. However, these traffic models are too specific for our constructions of quasi-output-buffered switches in this paper. Instead, we will use a much more general traffic characterization for flows of packets as in [12]. Throughout this paper, for a flow A, we denote $A(t)$ as the cumulative number of packets from flow A that arrive by time t for $t \geq 0$.

Definition 1 ***(λ-moment generating function bounded from above (λ-m.b.f.a.) flows)***

(i) A stochastic process $\{Q(t), t \geq 0\}$ is said to have a finite moment generating function *if there exists a real number $\theta > 0$ such that*

$$\sup_{t \geq 0} E[e^{\theta Q(t)}] < \infty. \tag{1}$$

(ii) We say that a flow A is λ-moment generating function bounded from above (λ-m.b.f.a.) *if the stochastic process $\{Q(t), t \geq 0\}$ defined in (2) below has a finite moment generating function for every $\epsilon > 0$:*

$$Q(t) = \max_{0 \leq s \leq t} [A(t) - A(s) - (\lambda + \epsilon)(t - s)]. \tag{2}$$

With $Q(0) = 0$, we note that $Q(t)$ in (2) is in fact the recursive expansion of the Lindley equation [14]

$$Q(t) = \max[0, Q(t-1) + a(t) - (\lambda + \epsilon)], \tag{3}$$

where $a(t) = A(t) - A(t-1)$ is the number of packets from flow A that arrive at time t. In view of (3), $Q(t)$ is simply the number of packets stored in the system at time t when we feed flow A to a work conserving link with capacity $\lambda + \epsilon$. It is known from the Loynes construction [15] that the stochastic process $\{Q(t), t \geq 0\}$ converges in distribution to a steady state random variable $Q(\infty)$ if the stochastic process $\{a(t), t \geq 1\}$ is stationary and ergodic with a mean rate not greater than λ.

However, traffic characterization by the mean rate of a stationary and ergodic process is not strong enough to guarantee that the steady state random variable $Q(\infty)$ has a finite moment generating function. For this, we need a stronger condition in [16]. For $\theta > 0$, let

$$a^*(\theta) = \limsup_{t \to \infty} \frac{1}{\theta t} \left[\sup_{s \geq 0} \log E[e^{\theta(A(t+s) - A(s))}] \right] \tag{4}$$

be the minimum envelope rate (MER) with respect to θ (note that $a^*(\theta)$ is also known as the effective bandwidth function in the literature, see e.g., [13]). For $\epsilon > 0$, let

$$Q_\theta(t) = \max_{0 \leq s \leq t} [A(t) - A(s) - (a^*(\theta) + \epsilon)(t - s)].$$

From Theorem 3.8 in [16], we know that

$$\sup_{t \geq 0} E[e^{\theta Q_\theta(t)}] < \infty.$$

Therefore, it follows from Definition 1 that flow A is $a^*(\theta)$-m.b.f.a. for any $\theta > 0$. One can further choose the best traffic characterization by letting $\rho = \inf_{\theta > 0} a^*(\theta)$ so that flow A is ρ-m.b.f.a.

We note that for many arrival processes, the value ρ is simply the "mean" arrival rate, as illustrated in the following example for the Bernoulli arrival process.

Example 2 *Consider the Bernoulli arrival process with mean arrival rate ρ, i.e., with probability ρ there is an arriving packet in a time slot and this is independent of everything else. For such an arrival process, it is easy to see that*

$$a^*(\theta) = \frac{1}{\theta} \log(\rho e^\theta + (1 - \rho)),$$

and

$$\inf_{\theta > 0} a^*(\theta) = \lim_{\theta \to 0} a^*(\theta) = \rho.$$

Therefore, the Bernoulli arrival process with mean arrival rate ρ is ρ-m.b.f.a.

In view of Example 2, our traffic characterization is only slightly stronger than the traffic characterization by the mean arrival rate. The additional assumption on the bounded moment generating functions leads to the following three important properties: the superposition property, the splitting property, and the departure property.

We first derive the superposition property for two flows in the following lemma.

Lemma 3 ***(Superposition)***

(i) If two stochastic processes $\{Q_1(t), t \geq 0\}$ and $\{Q_2(t), t \geq 0\}$ have finite moment generating functions, then the superposition $\{Q(t), t \geq 0\}$ of the two stochastic processes $\{Q_1(t), t \geq 0\}$ and $\{Q_2(t), t \geq 0\}$, defined by $Q(t) = Q_1(t) + Q_2(t)$ for $t \geq 0$, also has a finite moment generating function.

(ii) If flow A_1 is λ_1-m.b.f.a. and flow A_2 is λ_2-m.b.f.a., then the superposition $A_1 + A_2$ of the two flows A_1 and A_2, defined by $(A_1 + A_2)(t) = A_1(t) + A_2(t)$ for $t \geq 0$, is $(\lambda_1 + \lambda_2)$-m.b.f.a.

Proof. See Appendix A for a proof. ■

We note that the proof of Lemma 3 is based on Cauchy-Schwartz inequality, and the two stochastic processes $Q_1(t)$ and $Q_2(t)$ in Lemma 3(i) and the two flows A_1 and A_2 in Lemma 3(ii) need not be independent (see the proof of Lemma 3 in Appendix A for details). As discussed before, if we view λ_1 as the "mean" rate of flow A_1 and λ_2 as the "mean" rate of flow A_2, then the aggregated flow $A_1 + A_2$ has a "mean" rate equal to $\lambda_1 + \lambda_2$.

In the following lemma, we show the splitting property for a flow that is splitted into several subflows in a round-robin fashion.

Lemma 4 ***(Round-robin splitting)*** *Suppose that a flow A is splitted into p subflows $A_1, A_2, \ldots, A_p$ in a round-robin fashion such that*

$$A_m(t) = \left\lceil \frac{A(t) - m + 1}{p} \right\rceil, \ m = 1, 2, \ldots, p. \tag{5}$$

If flow A is λ-m.b.f.a., then subflow A_m is λ/p-m.b.f.a. for $m = 1, 2, \ldots, p$.

Proof. See Appendix B for a proof. ■

The intuition of Lemma 4 is quite obvious. If we view λ as the "mean" rate of flow A, then subflow A_m has a "mean" rate equal to λ/p as it is obtained from flow A via the round-robin splitting.

Finally, we give the departure property in the following lemma.

Lemma 5 ***(Departure)*** *Suppose that a flow A is fed into a system (possibly along with other flows) that is initially empty at time 0. Let flow B be the departure flow of flow A, namely, $B(t)$ is the cumulative number of packets from flow A that depart from the system by time t. Also, let $Q(t)$ be the total number of packets (including packets from flow A and other flows) stored in the system at time t. If flow A is λ-m.b.f.a. and $\{Q(t), t \geq 0\}$ has a finite moment generating function, then flow B is also λ-m.b.f.a.*

Proof. See Appendix C for a proof. ■

The departure property shows that if flow A has a "mean" rate equal to λ, then flow B, the departure flow of flow A, also has a "mean" rate equal to λ provided that the system is "stable" (in the sense that the total number of packets stored in the system has a finite moment generating function). We note that it is difficult to obtain the departure property in Lemma 5 if one uses weaker traffic characterizations, such as stationarity and ergodicity. On the other hand, it is possible to obtain such a departure property by using stronger traffic characterizations, such as the (σ, ρ)-deterministic traffic characterization in the network calculus [17]. However, such a deterministic traffic characterization cannot be used for stochastic analysis needed in our later development.

As we shall see later, the superposition property, the splitting property, and the departure property provide us with a simple calculus for our traffic characterization in a network of switches.

B. Output-Buffered Switches

A switch that has M input links and N output links is called an $M \times N$ switch. A (local) flow in an $M \times N$ switch is a sequence of packets that arrive from the same input link and destined for the same output link. As there are M input links and N output links, there are MN flows in an $M \times N$ switch.

Let flow $A_{i,j}$ be the flow from input link i to output link j and let $A_{i,j}(t)$ be the cumulative number of packets from flow $A_{i,j}$ that arrive by time t for $i = 1, 2, \ldots, M$ and $j = 1, 2, \ldots, N$. Let $B_j(t)$ be the cumulative number of packets that depart from output link j by time t for $j = 1, 2, \ldots, N$. In an $M \times N$ output-buffered switch as defined below, packets are stored at the output links and we let $Q_j(t)$ be the number of packets stored in the buffer at output link j at time t for $j = 1, 2, \ldots, N$.

Definition 6 ***(Output-buffered switches)*** *An $M \times N$ switch is called an $M \times N$* output-buffered switch *if it satisfies the following two properties when it is started from an empty system at time 0.*

(i) *Packets destined for the same output link depart in the FIFO order.*
(ii) *For $j = 1, 2, \ldots, N$, $Q_j(t)$ is given by*

$$Q_j(t) = \max \left[0, Q_j(t-1) + \sum_{i=1}^{M} a_{i,j}(t) - 1\right], \tag{6}$$

where $a_{i,j}(t) = A_{i,j}(t) - A_{i,j}(t-1)$ is the number of packets from flow $A_{i,j}$ that arrive at time t.

We note that the Lindley equation in (6) says that all of the packets that arrive at time t from flows $A_{1,j}, A_{2,j}, \ldots, A_{M,j}$ are sent to the buffer at output link j at the same time. If there are packets in the buffer at output link j at time t, then one packet will depart from output link j at time t. We note that there might be packets arriving from all of the M flows $A_{1,j}, A_{2,j}, \ldots, A_{M,j}$ at the same time in the worst case, and in that case the buffer at output link j is required to have the capability of receiving M packets at the same time. As such, each output buffer needs to speed up (at least) M times and that causes the notorious scalability problem for an output-buffered switch.

By recursively expanding the Lindley equation in (6) with $Q_j(0) = 0$ yields

$$Q_j(t) = \max_{0 \le s \le t} \left[\sum_{i=1}^{M} (A_{i,j}(t) - A_{i,j}(s)) - (t - s) \right], \qquad (7)$$

for $j = 1, 2, \ldots, N$. Since $Q_j(t) = \sum_{i=1}^{M} A_{i,j}(t) - B_j(t)$, it then follows that

$$B_j(t) = \min_{0 \le s \le t} \left[\sum_{i=1}^{M} A_{i,j}(s) + (t - s) \right], \qquad (8)$$

for $j = 1, 2, \ldots, N$. Note that from (8) and the FIFO property of an output-buffered switch, the departure of a packet at time t is uniquely determined by all of the packets that arrive by time t. Therefore, if the arrival times of all of the packets are delayed by c time slots, then the departure times of all of the packets are also delayed by c time slots.

To ensure the stability of an output-buffered switch, we need the following no overbooking condition.

Definition 7 ***(No overbooking condition)*** *The input traffic of an $M \times N$ switch is said to satisfy the* no overbooking condition *if flow $A_{i,j}$ is $\lambda_{i,j}$-m.b.f.a. for $i = 1, 2, \ldots, M$ and $j = 1, 2, \ldots, N$, and*

$$\sum_{i=1}^{M} \lambda_{i,j} < 1, \ \textit{for } j = 1, 2, \ldots, N. \qquad (9)$$

Intuitively, the no overbooking condition in (9) says that the total "mean" rate to a particular output link cannot exceed 1. Under the no overbooking condition, we show in Lemma 8 below that an output-buffered switch is stable in the sense that the total number of packets stored in the switch has a finite moment generating function. Such a stability property in Lemma 8 is called the *universal stability property*.

Lemma 8 ***(Universal stability)*** *Suppose that an $M \times N$ output-buffered switch is started from an empty system at time 0, and its input traffic satisfies the no overbooking condition in Definition 7. Then we have*

(i) $\{Q_j(t), t \ge 0\}$ has a finite moment generating function for $j = 1, 2, \ldots, N$.

(ii) Let $Q(t) = \sum_{j=1}^{N} Q_j(t)$ be the total number of packets stored in the switch at time t. Then $\{Q(t), t \ge 0\}$ has a finite moment generating function.

Proof. (i) As flow $A_{i,j}$ is $\lambda_{i,j}$-m.b.f.a. for $i = 1, 2, \ldots, M$ and $j = 1, 2, \ldots, N$, it follows from the superposition property in Lemma 3(ii) that the aggregated flow $\sum_{i=1}^{M} A_{i,j}$ to output link j is $\sum_{i=1}^{M} \lambda_{i,j}$-m.b.f.a. for $j = 1, 2, \ldots, N$.

Since $\sum_{i=1}^{M} \lambda_{i,j} < 1$ for $j = 1, 2, \ldots, N$, it then follows from (7) and Definition 1(ii) that $\{Q_j(t), t \ge 0\}$ has a finite moment generating function for $j = 1, 2, \ldots, N$.

(ii) As $Q(t) = \sum_{j=1}^{N} Q_j(t)$, it is clear from Lemma 8(i) and the superposition property in Lemma 3(i) that $\{Q(t), t \ge 0\}$ has a finite moment generating function. ■

C. Definition of Quasi-Output-Buffered Switches

As discussed before, output-buffered switches do not scale due to the needed speedup. As a result, it is difficult to construct a large output-buffered switch *directly*. A natural question is then whether one can construct a larger switch using a set of smaller switches. We will show in this paper that this is possible by extracting and preserving some key properties in output-buffered switches. The switches that satisfy these key properties are called quasi-output-buffered switches as defined in Definition 9 below, namely, they behave like output-buffered switches but they are not exactly the same as output-buffered switches.

Definition 9 ***(Quasi-output-buffered switches)*** *An $M \times N$ switch is called an $M \times N$* quasi-output-buffered switch *if it satisfies the following three properties when it is started from an empty system at time 0.*

(P1) *Deterministic mapping: The departure time of every packet is a deterministic function of the arrival times of all of the packets. This implies that if the arrival times of all of the packets are delayed by c time slots, then a quasi-output-buffered switch can be operated in such a way (by shifting the starting time of the switch) that the departure times of all of the packets are also delayed by c time slots.*

(P2) *FIFO delivery: Packets of the same flow depart in the FIFO order.*

(P3) *Universal stability: Let $Q(t)$ be the total number of packets stored in the switch at time t. If the input traffic of the switch satisfies the no overbooking condition in Definition 7, then $\{Q(t), t \ge 0\}$ has a finite moment generating function.*

From the discussions in Section II-B and Lemma 8, it is clear that an output-buffered switch is a quasi-output-buffered switch. It follows that the switches that achieve exact emulation of output-buffered switches (see e.g., [1]–[4]) are quasi-output-buffered switches. Various versions of load-balanced Birkhoff-von Neumann switches, including the Uniform Frame Spreading (UFS) in [6], the Padded Frame (PF) in [8], and the Contention and Reservation (CR) switch in [9], preserve the FIFO delivery of packets from the same flow and are shown to have a constant bound when compared to the total number of packets in the corresponding output-buffered switch. Thus, they are also quasi-output-buffered switches. However, it is not clear whether an input-buffered switch with maximum weight matching (MWM) [18] is a quasi-output-buffered switch as the universal stability property in (P3) of Definition 9 has not been proved in the literature yet. We also note that switches that use randomized algorithms (see e.g., [19]) are not quasi-output-buffered switches as they fail to satisfy the deterministic mapping property.

D. Feedforward Networks of Quasi-Output-Buffered Switches

In this section, we show that the key properties of quasi-output-buffered switches can be preserved through a feedforward network. To illustrate this, consider a feedforward network interconnected by K switches that are indexed from 1 to K. Suppose that the k^{th} switch in the feedforward network has I_k output links that are indexed from 1 to I_k for $k = 1, 2, \ldots, K$. Also suppose that there are F end-to-end flows that are indexed from 1 to F in the feedforward network. As the network is feedforward, we assume without loss of generality that every end-to-end flow traverses through the network in the increasing order of the indices of the switches. For $f = 1, 2, \ldots, F$, $j = 1, 2, \ldots, I_k$, and $k = 1, 2, \ldots, K$, let $D_{f,j}^{(k)}$ be the routing variable for the f^{th} end-to-end flow at output link j of the k^{th} switch, i.e., $D_{f,j}^{(k)} = 1$ if the f^{th} end-to-end flow traverses through output link j of the k^{th} switch and $D_{f,j}^{(k)} = 0$ otherwise.

Theorem 10 *Suppose that the K switches in the feedforward network are quasi-output-buffered switches and they are started from an empty system at time 0. Assume that the f^{th} end-to-end flow is λ_f-m.b.f.a. when it arrives at the feedforward network for $f = 1, 2, \ldots, F$, and assume that the total "mean" arrival rate at every output link of every quasi-output-buffered switch does not exceed its capacity, i.e.,*

$$\sum_{f=1}^{F} D_{f,j}^{(k)} \lambda_f < 1, \tag{10}$$

for $j = 1, 2, \ldots, I_k$ and $k = 1, 2, \ldots, K$.

(i) Let $Q^{(k)}(t)$ be the total number of packets stored in the k^{th} switch at time t for $k = 1, 2, \ldots, K$. Then $\{Q^{(k)}(t), t \geq 0\}$ has a finite moment generating function for $k = 1, 2, \ldots, K$.

(ii) The f^{th} end-to-end flow is λ_f-m.b.f.a. at every link traversed by the flow for $f = 1, 2, \ldots, F$.

(iii) Let $Q(t) = \sum_{k=1}^{K} Q^{(k)}(t)$ be the total number of packets stored in the feedforward network at time t. Then $\{Q(t), t \geq 0\}$ has a finite moment generating function.

Proof. Consider the first quasi-output-buffered switch. Note that there is only external traffic to the first switch. As $\sum_{f=1}^{F} D_{f,j}^{(1)} \lambda_f < 1$ for $j = 1, 2, \ldots, I_1$, the input traffic of the first switch satisfies the no overbooking condition, and it follows from the universal stability property in (P3) of Definition 9 that $\{Q^{(1)}(t), t \geq 0\}$ has a finite moment generating function. Assume that the f^{th} end-to-end flow traverses the first switch. Since the f^{th} end-to-end flow is λ_f-m.b.f.a. when it arrives at the first switch and we just showed that $\{Q^{(1)}(t), t \geq 0\}$ has a finite moment generating function, it then follows from the departure property in Lemma 5 that the f^{th} end-to-end flow is also λ_f-m.b.f.a. when it departs from the first switch. Therefore, (i) and (ii) hold for the first switch.

Now we consider the second quasi-output-buffered switch. The input traffic of the second switch is either external or from the output links of the first switch. As $\sum_{f=1}^{F} D_{f,j}^{(2)} \lambda_f < 1$ for $j = 1, 2, \ldots, I_2$, the input traffic of the second switch satisfies the no overbooking condition, and hence we can show that (i) and (ii) hold for the second switch by using the same argument for the first switch. It should be clear that by using $\sum_{f=1}^{F} D_{f,j}^{(k)} \lambda_f < 1$ for $j = 1, 2, \ldots, I_k$ and $k = 1, 2, \ldots, K$, and repeating the same argument for the first switch for K times, we can show that (i) and (ii) hold for the K switches in the feedforward network.

As $Q(t) = \sum_{k=1}^{K} Q^{(k)}(t)$, it is clear from Theorem 10(i) and the superposition property in Lemma 3(i) that $\{Q(t), t \geq 0\}$ has a finite moment generating function. ■

As shown in Theorem 10, the key condition in (10) is to make sure that the total "mean" arrival rate at every output link of every quasi-output-buffered switch does not exceed its capacity. This will be done by load balancing in the three-stage construction of a quasi-output-buffered switch in Section III.

III. A Three-Stage Construction of a Quasi-Output-Buffered Switch

A. Operation Rules for the Three-Stage Construction

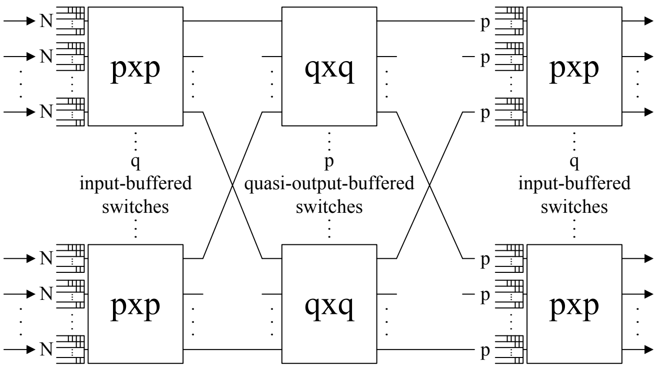


Fig. 1. A three-stage construction of an $N \times N$ quasi-output-buffered switch, where $N = p \times q$.

In this section, we show how one can construct a larger quasi-output-buffered switch by using a set of smaller quasi-output-buffered switches. In Figure 1, we show a three-stage construction of an $N \times N$ quasi-output-buffered switch, where $N = p \times q$. In the first stage, there are q $p \times p$ input-buffered switches. Each input buffer at an input link of a switch in the first stage has N virtual output queues (VOQ). In the second stage, there are p $q \times q$ quasi-output-buffered switches. Finally, in the third stage, there are also q $p \times p$ input-buffered switches. Each input buffer at an input link of a switch in the third stage has p VOQs. As in a standard Clos network [10], the switches in the first stage and those in the second stage are connected by the perfect shuffle exchange, i.e., for $m = 1, 2, \ldots, p$ and $\ell = 1, 2, \ldots, q$, output link m of the ℓ^{th} switch in the first stage is connected to input link ℓ of the m^{th} switch in the second stage. Similarly, the switches in the second stage and those in the third stage are also connected by the perfect shuffle exchange, i.e., for $m = 1, 2, \ldots, p$ and $\ell = 1, 2, \ldots, q$, output link ℓ of the m^{th} switch in the second stage is connected to input link m of the ℓ^{th} switch in the third stage.

The main idea of the three-stage construction is to accumulate packets in the first stage to form a frame of p packets. Then use the uniform frame spreading (UFS) scheme in [6] to distribute the packets in a frame *evenly* to the p quasi-output-buffered switches in the second stage. Finally, packets in a frame are "re-assembled" in the third stage.

To do this, the connection patterns of the $p \times p$ switches in the first stage and the third stage are specified by the $p \times p$

symmetric TDM switch in [20]. Recall that a $p \times p$ symmetric TDM switch implements the following periodic connection patterns: input link i of the $p \times p$ switch is connected to output link j of the $p \times p$ switch at time t if and only if

$$(i+j) \bmod p = (t+1) \bmod p. \tag{11}$$

In other words, for any positive integer f, input link i is connected to output link 1 at time $i+(f-1)p$, to output link 2 at time $i+(f-1)p+1, \ldots$, and to output link p at time $i+fp-1$. Also, it is clear from (11) that every connection pattern in a symmetric TDM switch is *symmetric* (as input link i is connected to output link j if and only if input link j is connected to output link i). As such, output link i is connected to input link 1 at time $i+(f-1)p$, to input link 2 at time $i+(f-1)p+1, \ldots$, and to input link p at time $i+fp-1$.

Now we specify the operation rules in these three stages.

(R1) Uniform frame spreading (UFS) for the symmetric TDM switches in the first stage: There are N VOQs at every input link of every $p \times p$ symmetric TDM switch in the first stage. When a packet destined for (external) output link j, where $1 \le j \le N$, arrives at an input link of a switch in the first stage, it is placed in the j^{th} VOQ at that input link. The switches in the first stage are operated in a frame-based manner as in the UFS scheme in [6]. Every frame consists of p consecutive time slots. However, the beginning time slots of frames are different for different *inputs*. Specifically, the f^{th} frame of input link i of a switch in the first stage begins at the f^{th} time when input link i of that switch is connected to the *first* output link of that switch, i.e., when input link i of that switch is connected to the first quasi-output-buffered switch in the second stage. As such, we have from the connection patterns of a $p \times p$ symmetric TDM switch in (11) that the f^{th} frame of input link i of a switch in the first stage consists of time slots $i+(f-1)p, i+(f-1)p+1, \ldots, i+fp-1$ for $i = 1, 2, \ldots, p$. If there are at least p packets in a VOQ at an input link of a switch in the first stage, then we call that VOQ a *full-framed VOQ*. Consider a switch in the first stage and consider a frame of an input link, say input link i, of the switch. If input link i of the switch has at least one full-framed VOQ at the beginning of the frame, then the switch selects one full-framed VOQ from input link i and sends the first p packets from the selected full-framed VOQ during the frame so that those p packets are distributed evenly to the p quasi-output-buffered switches in the second stage. Otherwise, the switch does nothing during the frame.

(R2) Time shifted operations for the quasi-output-buffered switches in the second stage: From the UFS scheme in the operation rule (R1), we know that if there is a packet destined for (external) output link j, where $1 \le j \le N$, that arrives at input link i of the *first* switch in the second stage at time t, then there is also a packet destined for output link j that arrives at input link i of the m^{th} switch in the second stage at time $t+m-1$ for $m = 2, 3, \ldots, p$. In other words, the arrival process to the m^{th} switch in the second stage is simply a time shifted version of that to the first switch in the second stage for $m = 2, 3, \ldots, p$. Therefore, they can be made to be *identical* if we run the clock in the m^{th} switch by the new time $t' = t - m + 1$ for $m = 2, 3, \ldots, p$. As it is easy to see that there is a *unique* routing path from an input link of a switch in the second stage to an (external) output link, it then follows from the deterministic mapping property in (P1) of Definition 9 that the departure process from the first switch in the second stage and that from the m^{th} switch in the second stage are also *identical* with respect to the new clock for $m = 2, 3, \ldots, p$. As such, if there is a packet destined for (external) output link j, where $1 \le j \le N$, that arrives at the *first* input of the $\lceil \frac{j}{p} \rceil^{\text{th}}$ switch in the third stage at time t (note that the $\lceil \frac{j}{p} \rceil^{\text{th}}$ switch in the third stage contains (external) output link j), then there is also a packet destined for output link j that arrives at input link m of the $\lceil \frac{j}{p} \rceil^{\text{th}}$ switch in the third stage at time $t+m-1$ for $m = 2, 3, \ldots, p$.

(R3) Inverse uniform frame spreading for the symmetric TDM switches in the third stage: There are p VOQs at every input link of every $p \times p$ symmetric TDM switch in the third stage. When a packet destined for (external) output link j, where $1 \le j \le N$, arrives at an input link of the $\lceil \frac{j}{p} \rceil^{\text{th}}$ switch in the third stage, it is placed in the $(j - (\lceil \frac{j}{p} \rceil - 1)p)^{\text{th}}$ VOQ at that input link. The switches in the third stage are operated in a frame-based manner as that for the switches in the first stage. Every frame consists of p consecutive time slots. However, the beginning time slots of frames are different for different *outputs*. Specifically, the f^{th} frame of output link i of a switch in the third stage begins at the f^{th} time when output link i of that switch is connected to the *first* input link of that switch, i.e., when output link i of that switch is connected to the first quasi-output-buffered switch in the second stage. As such, we have from the connection patterns of a $p \times p$ symmetric TDM switch in (11) that the f^{th} frame of output link i of a switch in the third stage consists of time slots $i+(f-1)p, i+(f-1)p+1, \ldots, i+fp-1$. Consider a switch in the third stage and consider a frame of an output link of the switch, say the f^{th} frame of output link i of the switch. From the time shifted operations in the operation rule (R2), we know that if the i^{th} VOQ of the *first* input link of the switch is not empty at time $i+(f-1)p$, then the i^{th} VOQ of input link m of the switch is also not empty at time $i+(f-1)p+m-1$ for $m = 2, 3, \ldots, p$, and every input link of the switch sends the head-of-line packet from its i^{th} VOQ to output link i of the switch during the frame. On the other hand, if the i^{th} VOQ of the first input link of the switch is empty at time $i+(f-1)p$, then the i^{th} VOQ of input link m of the switch is also empty at time $i+(f-1)p+m-1$ for $m = 2, 3, \ldots, p$, and the switch does nothing during the frame.

In the following theorem, we show the main result of this paper.

Theorem 11 *Suppose that the three-stage construction in Figure 1 is started from an empty system at time 0, and its input traffic satisfies the no overbooking condition in Definition 7 (with $M = N$), i.e., flow $A_{i,j}$ is $\lambda_{i,j}$-m.b.f.a. for $i = 1, 2, \ldots, N$ and $j = 1, 2, \ldots, N$, and*

$$\sum_{i=1}^{N} \lambda_{i,j} < 1, \text{ for } j = 1, 2, \ldots, N. \tag{12}$$

Then under the operation rules (R1)–(R3), the three-stage construction in Figure 1 can be operated as an $N \times N$ quasi-output-buffered switch.

We note there are several early works in the literature (see e.g., [21], [22]) that also used the three-stage Clos network to construct a larger switch. To the best of our knowledge, it seems that Theorem 11 on quasi-output-buffered switches is the first result that allows *recursive constructions* of switches with comparable performance (in the sense of 100% throughput and FIFO delivery of packets from the same flow) to output-buffered switches.

Clearly, as the switches in the first stage and the third stage are symmetric TDM switches, they are deterministic. As the quasi-output-buffered switches in the second stage satisfy the deterministic mapping property in (P1) of Definition 9, the three-stage construction in Figure 1 also satisfies the deterministic mapping property. Furthermore, as the quasi-output-buffered switches in the second stage satisfy the FIFO delivery property in (P2) of Definition 9, it then follows from the UFS scheme in (R1) and the inverse UFS scheme in (R3) that packets of the same flow depart in the FIFO order in the three-stage construction. Thus, the three-stage construction in Figure 1 also satisfies the FIFO delivery property. It remains to show the three-stage construction in Figure 1 satisfies the universal stability property in (P3) of Definition 9. This will be done in the following section.

B. Proof of the Universal Stability Property

In this section, we show the universal stability property for the three-stage construction in Section III-A.

Note that the switches in the first stage are operated under the UFS scheme. Using the arguments in [6] and [9], we show in the following proposition that the number of packets stored in an input buffer of a switch in the first stage in Figure 1 is bounded above by a finite constant.

Proposition 12 *The total number of packets stored in an input buffer of a switch in the first stage in Figure 1 is bounded above by $N(p-1)+p$.*

Proof. See Appendix D for a proof. ■

It is clear from Proposition 12 that the universal stability property is satisfied for the switches in the first stage. Now we show the universal stability property for the switches in the second stage. As the switches in the second stage are quasi-output-buffered switches, the key step is then to verify that the no overbooking condition is satisfied for every quasi-output-buffered switch in the second stage.

For $i = 1, 2, \ldots, N$ and $j = 1, 2, \ldots, N$, let flow $B_{i,j}^{(1)}$ be the departure flow of flow $A_{i,j}$ from the first stage. As flow $A_{i,j}$ is $\lambda_{i,j}$-m.b.f.a., it then follows from the departure property in Lemma 5 and Proposition 12 that flow $B_{i,j}^{(1)}$ is also $\lambda_{i,j}$-m.b.f.a.

For $i = 1, 2, \ldots, N$, $j = 1, 2, \ldots, N$, and $m = 1, 2, \ldots, p$, let $A_{i,j,m}^{(2)}(t)$ be the cumulative number of packets from flow $A_{i,j}$ that arrive at the m^{th} switch in the second stage by time t. As the switches in the first stage are operated under the UFS scheme in (R1), the packets from flow $B_{i,j}^{(1)}$ are distributed in a round-robin fashion to the switches in the second stage. Thus, we have

$$A_{i,j,m}^{(2)}(t) = \left\lceil \frac{B_{i,j}^{(1)}(t) - m + 1}{p} \right\rceil.$$

As flow $B_{i,j}^{(1)}$ is $\lambda_{i,j}$-m.b.f.a., it then follows from the splitting property in Lemma 4 that flow $A_{i,j,m}^{(2)}$ is $\lambda_{i,j}/p$-m.b.f.a.

For $k = 1, 2, \ldots, q$, $\ell = 1, 2, \ldots, q$, and $m = 1, 2, \ldots, p$, let flow $B_{k,\ell,m}^{(2)}$ be the local flow of packets that traverse from input link k of the m^{th} switch in the second stage to output link ℓ of that switch. Clearly, flow $B_{k,\ell,m}^{(2)}$ is the aggregated flow of the set of flows $A_{i,j,m}^{(2)}$, $i = (k-1)p+1, (k-1)p+2, \ldots, kp$, $j = (\ell-1)p+1, (\ell-1)p+2, \ldots, \ell p$. As flow $A_{i,j,m}^{(2)}$ is $\lambda_{i,j}/p$-m.b.f.a., we have from the superposition property in Lemma 3(ii) that the local flow $B_{k,\ell,m}^{(2)}$ is $(\sum_{i=(k-1)p+1}^{kp} \sum_{j=(\ell-1)p+1}^{\ell p} \lambda_{i,j}/p)$-m.b.f.a. From (12), we have

$$\begin{aligned}
&\sum_{k=1}^{q} \sum_{i=(k-1)p+1}^{kp} \sum_{j=(\ell-1)p+1}^{\ell p} \frac{\lambda_{i,j}}{p} \\
&= \sum_{i=1}^{N} \sum_{j=(\ell-1)p+1}^{\ell p} \frac{\lambda_{i,j}}{p} = \frac{1}{p} \sum_{j=(\ell-1)p+1}^{\ell p} \sum_{i=1}^{N} \lambda_{i,j} \\
&< \frac{1}{p} \sum_{j=(\ell-1)p+1}^{\ell p} 1 = 1,
\end{aligned}$$

for $\ell = 1, 2, \ldots, q$ and $m = 1, 2, \ldots, p$. As such, the no overbooking condition for the m^{th} quasi-output-buffered switch in the second stage is satisfied for $m = 1, 2, \ldots, p$. We then have the following proposition on the universal stability property for the switches in the second stage.

Proposition 13 *(i) Let $Q_m^{(2)}(t)$ be the total number of packets stored in the m^{th} quasi-output-buffered switch in the second stage at time t for $m = 1, 2, \ldots, p$. Then $\{Q_m^{(2)}(t), t \geq 0\}$ has a finite moment generating function for $m = 1, 2, \ldots, p$.*

(ii) Let $Q^{(2)}(t) = \sum_{m=1}^{p} Q_m^{(2)}(t)$ be the total number of packets stored in the second stage at time t. Then $\{Q^{(2)}(t), t \geq 0\}$ also has a finite moment generating function.

Proof. (i) As the no overbooking condition for the m^{th} quasi-output-buffered switch in the second stage is satisfied for $m = 1, 2, \ldots, p$, it follows from the universal stability property in (P3) of Definition 9 that $\{Q_m^{(2)}(t), t \geq 0\}$ has a finite moment generating function for $m = 1, 2, \ldots, p$.

(ii) As $Q^{(2)}(t) = \sum_{m=1}^{p} Q_m^{(2)}(t)$, it is clear from Proposition 13(i) and the superposition property in Lemma 3(i) that $\{Q^{(2)}(t), t \geq 0\}$ has a finite moment generating function. ■

Now we show the universal stability property for the switches in the third stage. For $i = 1, 2, \ldots, N$, $j = 1, 2, \ldots, N$, and $m = 1, 2, \ldots, p$, let $A_{i,j,m}^{(3)}(t)$ be the cumulative number of packets from flow $A_{i,j}$ that arrive at input link m of the $\lceil \frac{j}{p} \rceil^{\text{th}}$ switch in the third stage by time t (note that the $\lceil \frac{j}{p} \rceil^{\text{th}}$ switch in the third stage contains (external) output link j). Since flow $A^{(3)}_{i,j,m}$ is simply the departure flow of flow $A^{(2)}_{i,j,m}$ and flow $A^{(2)}_{i,j,m}$ is $\lambda_{i,j}/p$-m.b.f.a., we have from the departure property in Lemma 5 and Proposition 13(ii) that flow $A^{(3)}_{i,j,m}$ is also $\lambda_{i,j}/p$-m.b.f.a. For $j = 1, 2, \ldots, N$ and $m = 1, 2, \ldots, p$, let flow $A^{(3)}_{j,m}$ be the aggregated flow of the set of flows $A^{(3)}_{i,j,m}$, $i = 1, 2, \ldots, N$. Then we have from the superposition property in Lemma 3(ii) that the aggregated flow $A^{(3)}_{j,m}$ is $(\sum_{i=1}^{N} \lambda_{i,j}/p)$-m.b.f.a.

For $j = 1, 2, \ldots, N$ and $m = 1, 2, \ldots, p$, let $Q^{(3)}_{j,m}(t)$ be the total number of packets destined for (external) output j that are stored in the mth input buffer of the $\lceil \frac{j}{p} \rceil$th switch in the third stage at time t, and let $C^{(3)}_{j,m}(t)$ be the cumulative number of time slots that input link m of that switch is connected to (external) output j by time t. As the connection pattern of the switches in the third stage is periodic with period p, we have

$$C^{(3)}_{j,m}(t) - C^{(3)}_{j,m}(s) \geq \left\lfloor \frac{t-s}{p} \right\rfloor > \frac{t-s}{p} - 1.$$

Moreover, we have from the Lindley equation (with $Q^{(3)}_{j,m}(0) = 0$) that

$$\begin{aligned} Q^{(3)}_{j,m}(t) &= \max[0, Q^{(3)}_{j,m}(t-1) + A^{(3)}_{j,m}(t) - A^{(3)}_{j,m}(t-1) \\ &\qquad -(C^{(3)}_{j,m}(t) - C^{(3)}_{j,m}(t-1))] \\ &= \max_{0 \leq s \leq t} [A^{(3)}_{j,m}(t) - A^{(3)}_{j,m}(s) - (C^{(3)}_{j,m}(t) - C^{(3)}_{j,m}(s))] \\ &\leq \max_{0 \leq s \leq t} \left[A^{(3)}_{j,m}(t) - A^{(3)}_{j,m}(s) - \frac{1}{p}(t-s) \right] + 1. \end{aligned}$$

Since the aggregated flow $A^{(3)}_{j,m}$ is $\sum_{i=1}^{N} \lambda_{i,j}/p$-m.b.f.a. and we have from the no overbooking condition in (12) that $\sum_{i=1}^{N} \lambda_{i,j}/p < 1/p$, it then follows from Definition 1 that $\{Q^{(3)}_{j,m}(t), t \geq 0\}$ has a finite moment generating function. Using the superposition property in Lemma 3(i), we then have the following proposition on the universal stability property for the switches in the third stage.

Proposition 14 *Let $Q^{(3)}(t) = \sum_{j=1}^{N} \sum_{m=1}^{p} Q^{(3)}_{j,m}(t)$ be the total number of packets in the third stage at time t. Then $\{Q^{(3)}(t), t \geq 0\}$ has a finite moment generating function.*

Finally, let $Q(t)$ be the total number of packets inside the three-stage construction at time t. From Proposition 12, Proposition 13, Proposition 14, and the superposition property in Lemma 3(i), we then conclude that $\{Q(t), t \geq 0\}$ also has a finite moment generating function. Therefore, the universal stability property in (P3) of Definition 9 is satisfied for the three-stage construction in Figure 1.

IV. Packet-Pair Switches

A. Packet-Pair Switches Via Recursive Expansions of the Three-Stage Construction

In this section, we assume that N is a power of 2. In this case, we can recursively construct an $N \times N$ quasi-output-buffered switch by the three-stage construction in Figure 1 (as in the construction of a Benes network [11]). To do this, we first note that for $N = 2$ we can simply choose $p = 2$ and $q = 1$ in the three-stage construction in Figure 1. Since a 1×1 switch can be simply replaced by a single link, the three-stage construction for this is equivalent to the (two-stage) load-balanced Birkhoff-von Neumann switch with the UFS scheme. For such a switch, the frame size is 2 and packets are transmitted in pairs under the UFS scheme. Now we define packet-pair switches recursively as follows.

Definition 15 (Packet-pair switches)
(i) A 2×2 packet-pair switch is the 2×2 load-balanced Birkhoff-von Neumann switch with the UFS scheme.

(ii) An $N \times N$ packet-pair switch is constructed by the three-stage construction in Figure 1 with $p = 2$ and $q = N/2$, i.e., there are $N/2$ 2×2 input-buffered switches in the first stage, two $\frac{N}{2} \times \frac{N}{2}$ packet-pair switches in the second stage, and $N/2$ 2×2 input-buffered switches in the third stage.

By recursively expanding the $N \times N$ packet-pair switch, we have a multistage network of $2 \log_2 N$ stages with each stage consisting of $N/2$ 2×2 switches. In Figure 2, we show an 8×8 packet-pair switch.

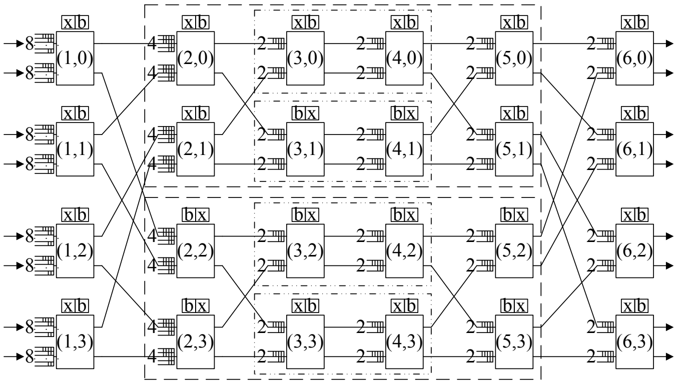


Fig. 2. An 8×8 packet-pair switch.

The operations of a packet-pair switch can also be specified in details by recursively expanding the operations in (R1) and (R3). In rules (R4) and (R5) below, we describe the detailed operations of an $N \times N$ packet-pair switch with $N = 2^n$. Note that for ease of presentation of rules (R4) and (R5), we index the (external) input links and (external) output links from 0 to $2^n - 1$ (instead of from 1 to 2^n). Also, the $N/2$ switches at each stage are indexed from 0 to $2^{n-1} - 1$ (instead of from 1 to 2^{n-1}).

(R4) Uniform frame spreading (UFS) for the first n stages: For $k = 1, 2, \ldots, n$, there are 2^{n-k+1} VOQs, indexed from 0 to $2^{n-k+1} - 1$, at every input link of every 2×2 switch in the kth stage. For $k = 1, 2, \ldots, n$ and $m = 0, 1, \ldots, 2^{n-1} - 1$, the connection patterns of the mth 2×2 switch in the kth stage are periodic with period 2. It is set to the "bar" state when

$$t + \sum_{\ell=2}^{k} \left\lfloor \frac{m \bmod 2^{n-\ell+1}}{2^{n-\ell}} \right\rfloor$$

is an odd number and to the "cross" state otherwise. Suppose that a packet destined for (external) output link j arrives at an input link of a switch in the k^{th} stage, where $0 \leq j \leq 2^n - 1$ and $1 \leq k \leq n$. Let $b_n b_{n-1} \dots b_1$ be the binary representation for j, i.e., $j = \sum_{\ell=1}^{n} b_\ell 2^{\ell-1}$. Then the packet is routed to the j_k^{th} VOQ of that input link, where $j_k = \sum_{\ell=1}^{n-k+1} b_{\ell+k-1} 2^{\ell-1}$. For $k = 1, 2, \dots, n$, a VOQ at an input link of a switch in the k^{th} stage is called a full-framed VOQ if there are at least two packets in that VOQ. When an input link of a switch is connected to output link 0 of that switch at time t, if there are full-framed VOQs at that input link, then the switch selects a full-framed VOQ from that input link and sends the first two packets (packet-pair) from the selected full-framed VOQ at time t and $t+1$. Otherwise, the switch does nothing at time t and $t+1$.

(R5) Self-routing for the last n stages: For $k = n+1, n+2, \dots, 2n$, there are two VOQs, indexed by 0 and 1, at every input link of every 2×2 switch in the k^{th} stage. For $k = n+1, n+2, \dots, 2n$ and $m = 0, 1, \dots, 2^{n-1} - 1$, the connection patterns of the m^{th} 2×2 switch in the k^{th} stage are the same as those of the m^{th} switch in the $(2n+1-k)^{\text{th}}$ stage. Suppose that a packet destined for (external) output link j arrives at an input link of a switch in the k^{th} stage, where $0 \leq j \leq 2^n - 1$ and $n+1 \leq k \leq 2n$. Let $b_n b_{n-1} \dots b_1$ be the binary representation for j, i.e., $j = \sum_{\ell=1}^{n} b_\ell 2^{\ell-1}$. Then the packet is routed to the j_k^{th} VOQ of that input link, where $j_k = b_{2n-k+1}$. When a switch is in the "bar" state at time t, VOQ 0 (resp., VOQ 1) at input link 0 (resp., input link 1) of the switch is selected and its head-of-line packet (if any) is transmitted at time t. Otherwise, VOQ 1 (resp., VOQ 0) at input link 0 (resp., input link 1) of the switch is selected and its head-of-line packet (if any) is transmitted at time t.

Note that the 2×2 switches in the first n stages of the $N \times N$ packet-pair switch is operated under the UFS scheme with frame size 2. From Proposition 12 (with $p = 2$), we know that the total number of packets in an input buffer of a switch in the k^{th} stage is bounded above by $2^{n-k+1} + 2$ for $k = 1, 2, \dots, n$. Moreover, we have from the deterministic mapping property that the arrival process to *any* input link of a 2×2 switch in the $(n+1)^{\text{th}}$ stage is simply a time shifted version of the arrival process to input link 0 of the 0^{th} switch in the $(n+1)^{\text{th}}$ stage. In view of this, the first n stages indeed perform load balancing for the incoming traffic at the $N \times N$ packet-pair switch. We illustrate this by the 8×8 packet-pair switch in Figure 2. As the first three stages are operated under the UFS scheme, the arrival process to switch (2,2) (resp., (2,3), (3,1), (3,3)) is one time slot later than that to switch (2,0) (resp., (2,1) (3,0), (3,2)). As the operations are deterministic, the departure process from switch (2,2) (resp., (2,3), (3,1), (3,3)) is also one time slot later than that from switch (2,0) (resp., (2,1) (3,0), (3,2)). As such, the arrival process to switch (3,2) is one time slot later than that to switch (3,0), and hence the arrival process to switch (3,3) is two time slots later than that to switch (3,0). As the arrival process to switch $(4, m)$ is simply the departure process from switch (3,m) for $m = 0, 1, 2, 3$, we then conclude that the arrival process to switch (4,2) is one time slot later than that to switch (4,0), and the arrival process to switch (4,3) is two time slots later than that to switch (4,0).

Now we consider the Bernoulli arrival traffic in Example 2. With probability $0 \leq \rho < 1$, there is a packet arriving at an input link of the $N \times N$ packet-pair switch. This is independent of everything else. With probability $r_{i,j}$, an arriving packet at input link i is destined for output link j for $i = 1, 2, \dots, N$ and $j = 1, 2, \dots, N$. This is also independent of everything else. For such a model, flow $A_{i,j}$ (the sequence of packets from input link i to output link j) is a Bernoulli arrival process with mean $\rho r_{i,j}$ for $i = 1, 2, \dots, N$ and $j = 1, 2, \dots, N$. From Example 2, we know that flow $A_{i,j}$ is $\lambda_{i,j}$-m.b.f.a., where $\lambda_{i,j} = \rho r_{i,j}$, for $i = 1, 2, \dots, N$ and $j = 1, 2, \dots, N$. To ensure the universal stability, we assume the following no overbooking condition:

$$\sum_{i=1}^{N} \lambda_{i,j} < 1, \text{ for } j = 1, 2, \dots, N, \tag{13}$$

or equivalently,

$$\sum_{i=1}^{N} r_{i,j} < \frac{1}{\rho}, \text{ for } j = 1, 2, \dots, N. \tag{14}$$

As the $N \times N$ packet-pair switch is a quasi-output-buffered switch, the next theorem then follows from the no overbooking condition in (13) and the universal stability property in (P3) of Definition 9.

Theorem 16 *For the Bernoulli arrival traffic described above, there exists a $\theta > 0$ such that*

$$\sup_{t \geq 0} E[e^{\theta Q(t)}] < \infty, \tag{15}$$

where $Q(t)$ is the total number of packets stored in the $N \times N$ packet-pair switch at time t.

In summary, the packet-pair switches have the following nice features:

1) They achieve 100% throughput.
2) They deliver packets of the same flow in the FIFO order.
3) They only contain 2×2 switches and the connection patterns of these 2×2 switches are deterministic and periodic with period 2.
4) Packets are self-routed through the network of 2×2 switches.
5) No computation and communication is needed.

We note that the idea of using uniform traffic spreading and self routing in a buffered Benes network was previously used in [23], [24]. However, there is no guarantee that packets from the same flow are delivered in the FIFO order in [23], [24].

B. Delay Analysis of Packet-Pair Switches

To gain some intuition on the delay performance of the packet-pair switches, let us consider the *uniform* Bernoulli traffic, i.e., $r_{i,j} = 1/N$ for $i = 1, 2, \dots, N$ and $j = 1, 2, \dots, N$.

For a 2×2 switch in the *first* stage, there are N VOQs at every input link of the 2×2 switch. Recall that the operation of a 2×2 switch in the first stage is to transmit packets from a full-framed VOQ (if any) at an input link of the 2×2 switch when that input link is connected to the first output link of the 2×2 switch. A full-framed VOQ in this case is simply a VOQ that contains at least two packets. As such, we can implement the N VOQs at an input link of a 2×2 switch in the first stage by two parts: the first part is for storing packets that have not been "paired," and the second part is for storing packets that have been "paired." For this, there are N queues, indexed from 1 to N, with buffer size 1 in the first part, and there are two VOQs (for the two output links of the 2×2 switch) in the second part. Suppose that a packet destined for (external) output link j, where $1 \leq j \leq N$, arrives at an input of a 2×2 switch in the first stage. If the j^{th} queue in the first part is empty, then the arriving packet is placed in the j^{th} queue. On the other hand, if the j^{th} queue in the first part is not empty, then the arriving packet and the packet stored in the j^{th} queue are "paired" and they are moved to one of the two VOQs in the second part.

In view of the two-part implementation of the N VOQs at every input link of every 2×2 switch in the first stage, the delay at a 2×2 switch in the first stage consists of two parts: (i) the delay for "pairing" and (ii) the queueing delay for transmitting through the 2×2 switch. To compute the pairing delay, note that only the "odd" numbered packets in a flow need to wait for pairing, and the pairing delay for an odd numbered packet is simply the interarrival time of the next packet. Under the uniform Bernoulli traffic, the expected interarrival time of packets in every flow is N/ρ. Thus, the expected pairing delay is $N/2\rho$. For the queueing delay, we approximate the arrival process to the two VOQs in the second part by the Bernoulli arrival traffic with arrival rate ρ. As the connection pattern of every 2×2 switch is periodic with period 2, this model is a special case of the uniform Bernoulli traffic model in [5] (with $N = 2$). Thus, the expected queueing delay can be approximated by $1/2(1-\rho)$. Adding these two parts of delay, the expected delay through a 2×2 switch in the first stage can be approximated by

$$\frac{N}{2\rho} + \frac{1}{2(1-\rho)}.$$

If we approximate the arrival process to every input of every 2×2 switch in the packet-pair switch by the uniform Bernoulli traffic with arrival rate ρ, then using the same argument as that for the first stage yields the following approximation for the expected delay through a 2×2 switch in the k^{th} stage:

$$\frac{N}{2^k \rho} + \frac{1}{2(1-\rho)}, \ k = 1, 2, \ldots, n, \tag{16}$$

$$\frac{1}{2(1-\rho)}, \ k = n+1, n+2, \ldots, 2n, \tag{17}$$

where $n = \log_2 N$, as there is no "pairing" delay for the last n stages. Summing up the delays in (16) and (17), we can approximate the expected delay through the $N \times N$ packet-pair switch by

$$\frac{N-1}{\rho} + \frac{\log_2 N}{(1-\rho)}. \tag{18}$$

In Figure **??**, we compare our approximation in (18) with computer simulations. As shown in Figure **??**, our approximation (APPR) is a conservative estimate of the delay of the packet-pair (PP) switch. The reason for that is the arrival process to every input of a 2×2 switch in the packet-pair switch is *not* the uniform Bernoulli traffic. In fact, it is much more regular (less random) than the uniform Bernoulli traffic. This is because "pairing" takes time and it is less likely to have two consecutive pairs with the same destination.

To reduce the "pairing" delay of the packet-pair switch in light traffic, we can use the idea proposed in the padded frame (PF) scheme in [8]. At the beginning of a frame, if there are no full-framed VOQs in an input buffer of a switch in the first n stages, we can pad a fake packet to a VOQ with only one packet to form a padded frame (with frame size 2). Then the padded frame is transmitted inside the packet-pair switch. Clearly, it is most beneficial to generate padded frames in the first stage. The gain starts to diminish as the number of stages is increased. For this, we define a parameter n^+ as the number of stages that allow padded frames to be generated. To ensure stability, the number of padded frames inside the packet-pair switch has to be restrained. For this, we only allow padded frames to be generated when the total number of packets in the *first* input buffer of the *first* switch in the $(n+1)^{\text{th}}$ stage does not exceed a threshold TH. Such an enhancement is called a packet-pair-plus (PP$^+$) switch in this paper.

C. Simulations

In this section, we perform various simulations for packet-pair switches. In all of our simulations, the switch size N is chosen to be 64. Each simulation run contains 10^6 time slots. In Figure 3, we consider the uniform Bernoulli traffic model and plot the delays of the uniform frame spreading (UFS) scheme in [6], the padded frame (PF) scheme in [8], the Contention and Reservation (CR) switch in [9], the packet-pair (PP) switch, the packet-pair-plus (PP$^+$) switch, and the ideal output-buffered switch (OB). Certainly, the output-buffered switch has the best delay performance (at the cost of N times speedup). The packet-pair switch outperforms both the UFS scheme and the PF scheme. It also beats the CR switch in heavy traffic. However, its delay is higher than that in the CR switch in light traffic. This is because the CR switch uses the contention mode in light traffic, while the packet-pair switch wastes a lot of time to form a frame of two packets in light traffic. In this simulation, the packet-pair-plus switch is run with $n^+ = 3$ and $TH = 2$, i.e., only the first 3 stages are allowed to generate padded frames when the total number of packets in the first input buffer of the first switch in the 7th stage does not exceed 2. The delay of the PP^+ switch is much better than that of the PP switch in light traffic and is comparable to that of the PP switch in heavy traffic. Similar results are also shown in Figure 4 for the uniform Pareto traffic model in [5].

V. Conclusion

In this paper, we proposed a new concept, called *quasi-output-buffered switch*. Like an output-buffered switch, a quasi-output-buffered switch is a deterministic switch that achieves 100% throughput and delivers packets from the same flow in the FIFO order. Using the three-stage Clos network, we showed that one can *recursively* construct a larger quasi-output-buffered switch with a set of smaller quasi-output-buffered switches. By recursively expanding the three-stage

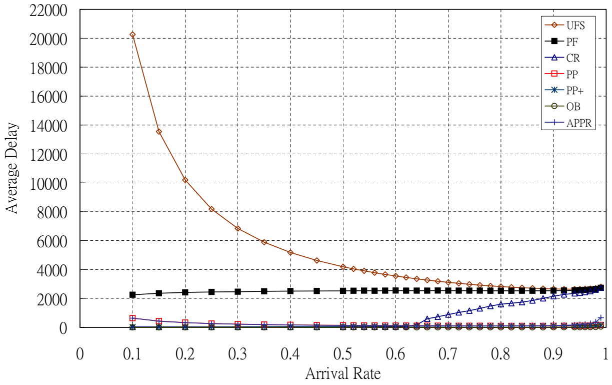


Fig. 3. Delay comparison for the uniform Bernoulli traffic model.

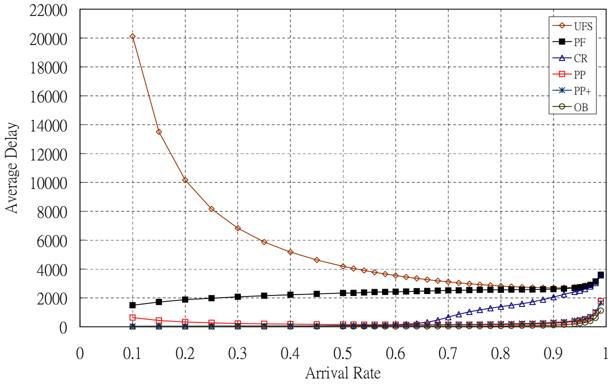


Fig. 4. Delay comparison for the uniform Pareto traffic model.

network, we obtained a packet-pair switch with only 2×2 switches. By computer simulations, we showed that packet-pair switches have better delay performance than most load-balanced switches with comparable construction complexity.

There are several problems that require further study:

(i) As argued in (18), the $N \times N$ packet-pair switch has $O(N)$ delay. It is shown in [25] that it is possible to obtain $O(\log N)$ delay in an $N \times N$ input-buffered switch (though this is at the cost of non-scalable computation and communication overheads by using Birkhoff-von Neumann decomposition). It would be of interest to find a scalable switch architecture that achieves $O(\log N)$ delay without any computation and communication.

(ii) We note that it is possible to replace the deterministic 2×2 switches in a packet-pair switch by fixed interconnecting networks. As such, one might be able to embed a packet-pair switch inside a fixed interconnecting network, e.g., a DWDM network. The problem is then how to do this efficiently.

(iii) A key distinguishing feature of output-buffered switches lies in the ability to control the departure order of packets from the switches to achieve performance (e.g., delay and bandwidth) guarantees for different traffic flows [26]–[29]. A future direction of investigation along this line is to devise scheduling schemes for the quasi-output-buffered switches proposed in this paper so that they are capable of providing quality-of-service (QoS) guarantees for different traffic flows.

Appendix A
Proof of Lemma 3

(i) Since both $\{Q_1(t), t \ge 0\}$ and $\{Q_2(t), t \ge 0\}$ have finite moment generating functions, there exist $\theta_1 > 0$ and $\theta_2 > 0$ such that

$$\sup_{t \ge 0} E[e^{\theta_i Q_i(t)}] < \infty, \ \ i = 1, 2. \tag{19}$$

Let $\theta = \min[\theta_1, \theta_2]/2$. It then follows from $Q(t) = Q_1(t) + Q_2(t)$ for $t \ge 0$, Cauchy-Schwartz inequality, and (19) that

$$\begin{aligned}\sup_{t \ge 0} E[e^{\theta Q(t)}] &= \sup_{t \ge 0} E[e^{\theta (Q_1(t)+Q_2(t))}] \\ &\le \sup_{t \ge 0} \Big(E[e^{2\theta Q_1(t)}]\Big)^{1/2} \Big(E[e^{2\theta Q_2(t)}]\Big)^{1/2} \\ &\le \sup_{t \ge 0} \Big(E[e^{\theta_1 Q_1(t)}] E[e^{\theta_2 Q_2(t)}]\Big)^{1/2} \\ &\le \Big(\sup_{t \ge 0} E[e^{\theta_1 Q_1(t)}] \sup_{t \ge 0} E[e^{\theta_2 Q_2(t)}]\Big)^{1/2} \\ &< \infty.\end{aligned}$$

Therefore, $\{Q(t), t \ge 0\}$ also has a finite moment generating function.

(ii) For $\epsilon > 0$, let

$$\begin{aligned}Q(t) = \max_{0 \le s \le t} [&(A_1(t) + A_2(t)) - (A_1(s) + A_2(s)) \\ &-(\lambda_1 + \lambda_2 + \epsilon)(t - s)].\end{aligned}$$

Note that

$$\begin{aligned}Q(t) &= \max_{0 \le s \le t} \Big[A_1(t) - A_1(s) - \Big(\lambda_1 + \frac{\epsilon}{2}\Big)(t - s) \\ &\qquad +A_2(t) - A_2(s) - \Big(\lambda_2 + \frac{\epsilon}{2}\Big)(t - s)\Big] \\ &\le \max_{0 \le s \le t} \Big[A_1(t) - A_1(s) - \Big(\lambda_1 + \frac{\epsilon}{2}\Big)(t - s)\Big] \\ &\quad + \max_{0 \le s \le t} \Big[A_2(t) - A_2(s) - \Big(\lambda_2 + \frac{\epsilon}{2}\Big)(t - s)\Big] \\ &= Q_1(t) + Q_2(t),\end{aligned} \tag{20}$$

where

$$\begin{aligned}Q_1(t) &= \max_{0 \le s \le t} \Big[A_1(t) - A_1(s) - \Big(\lambda_1 + \frac{\epsilon}{2}\Big)(t - s)\Big], \\ Q_2(t) &= \max_{0 \le s \le t} \Big[A_2(t) - A_2(s) - \Big(\lambda_2 + \frac{\epsilon}{2}\Big)(t - s)\Big].\end{aligned}$$

Since flow A_1 is λ_1-m.b.f.a. and flow A_2 is λ_2-m.b.f.a., we see from Definition 1 that both $\{Q_1(t), t \ge 0\}$ and $\{Q_2(t), t \ge 0\}$ have finite moment generating functions. It then follows from (20) and Lemma 3(i) that $\{Q(t), t \ge 0\}$ also has a finite moment generating function. Therefore, the aggregated flow $A_1 + A_2$ of the two flows A_1 and A_2 is $(\lambda_1 + \lambda_2)$-m.b.f.a.

Appendix B
Proof of Lemma 4

From (5), we have

$$\begin{aligned}A_m(t) - A_m(s) &= \left\lceil \frac{A(t) - m + 1}{p} \right\rceil - \left\lceil \frac{A(s) - m + 1}{p} \right\rceil \\ &\le \left\lceil \frac{A(t) - A(s)}{p} \right\rceil \\ &< \frac{A(t) - A(s)}{p} + 1\end{aligned} \tag{21}$$

Since flow A is λ-m.b.f.a., we see from Definition 1 that for every $p\epsilon > 0$ there exists a $\theta/p > 0$ such that

$$\sup_{t\geq 0} E\left[e^{\frac{\theta}{p}\max_{0\leq s\leq t}[A(t)-A(s)-(\lambda+p\epsilon)(t-s)]}\right] < \infty. \tag{22}$$

It follows from (21) and (22) that

$$\begin{aligned}
&\sup_{t\geq 0} E\left[e^{\theta\max_{0\leq s\leq t}[A_m(t)-A_m(s)-(\lambda/p+\epsilon)(t-s)]}\right] \\
&\leq \sup_{t\geq 0} E\left[e^{\frac{\theta}{p}\max_{0\leq s\leq t}[A(t)-A(s)+p-(\lambda+p\epsilon)(t-s)]}\right] \\
&= \sup_{t\geq 0} e^{\theta} E\left[e^{\frac{\theta}{p}\max_{0\leq s\leq t}[A(t)-A(s)-(\lambda+p\epsilon)(t-s)]}\right] \\
&< \infty.
\end{aligned}$$

Therefore, the subflow A_m is λ/p-m.b.f.a. for $m = 1, 2, \ldots, p$.

Appendix C
Proof of Lemma 5

Since the system is initially empty at time 0 and packets cannot depart from the system before they arrive at the system (causality), it is clear that

$$B(t) \leq A(t). \tag{23}$$

For packets that have arrived by time t, either they are stored in the system at time t or they have departed from the system by time t. As $Q(t)$ is the total number of packets (including packets from flow A and other flows) stored in the system at time t, it is also clear that

$$B(t) + Q(t) \geq A(t). \tag{24}$$

From (23) and (24), we have

$$\begin{aligned}
&\max_{0\leq s\leq t}[B(t)-B(s)-(\lambda+\epsilon)(t-s)] \\
&\leq \max_{0\leq s\leq t}[A(t)-A(s)+Q(s)-(\lambda+\epsilon)(t-s)].
\end{aligned} \tag{25}$$

As $\{Q(t), t \geq 0\}$ has a finite moment generating function, there exists a $\theta_1 > 0$ such that

$$c_1 = \sup_{t\geq 0} E[e^{\theta_1 Q(t)}] < \infty. \tag{26}$$

Since flow A is λ-m.b.f.a., we see from Definition 1 that for every $\epsilon/2 > 0$ there exists a $\theta_2 > 0$ such that

$$c_2 = \sup_{t\geq 0} E[e^{\theta_2 \max_{0\leq s\leq t}[A(t)-A(s)-(\lambda+\epsilon/2)(t-s)]}] < \infty. \tag{27}$$

Let $\theta = \min[\theta_1, \theta_2]/2$. Then we have from (25) that

$$\begin{aligned}
&E[e^{\theta\max_{0\leq s\leq t}[B(t)-B(s)-(\lambda+\epsilon)(t-s)]}] \\
&\leq E[e^{\theta\max_{0\leq s\leq t}[A(t)-A(s)+Q(s)-(\lambda+\epsilon)(t-s)]}] \\
&= \max_{0\leq s\leq t} E[e^{\theta(A(t)-A(s)+Q(s)-(\lambda+\epsilon)(t-s))}] \\
&\leq \sum_{s=0}^{t} E[e^{\theta(A(t)-A(s)+Q(s)-(\lambda+\epsilon)(t-s))}] \\
&= \sum_{s=0}^{t} e^{-\theta\epsilon(t-s)/2} E[e^{\theta Q(s)} e^{\theta(A(t)-A(s)-(\lambda+\epsilon/2)(t-s))}].
\end{aligned} \tag{28}$$

From Cauchy-Schwartz inequality, $\theta = \min[\theta_1, \theta_2]/2$, (26), and (27), we have

$$\begin{aligned}
&E[e^{\theta Q(s)} e^{\theta(A(t)-A(s)-(\lambda+\epsilon/2)(t-s))}] \\
&\leq \left(E[e^{2\theta Q(s)}]\right)^{1/2} \left(E[e^{2\theta(A(t)-A(s)-(\lambda+\epsilon/2)(t-s))}]\right)^{1/2} \\
&\leq \left(E[e^{\theta_1 Q(s)}]\right)^{1/2} \left(E[e^{\theta_2(A(t)-A(s)-(\lambda+\epsilon/2)(t-s))}]\right)^{1/2} \\
&\leq \sqrt{c_1 c_2}.
\end{aligned} \tag{29}$$

As such, we have from (28) and (29) that

$$\begin{aligned}
&\sup_{t\geq 0} E[e^{\theta\max_{0\leq s\leq t}[B(t)-B(s)-(\lambda+\epsilon)(t-s)]}] \\
&\leq \sup_{t\geq 0} \sum_{s=0}^{t} e^{-\theta\epsilon(t-s)/2}\sqrt{c_1c_2} \\
&= \sup_{t\geq 0} \sum_{s=0}^{t} e^{-\theta\epsilon s/2}\sqrt{c_1c_2} \\
&= \sqrt{c_1c_2}\sum_{s=0}^{\infty} e^{-\theta\epsilon s/2} = \frac{\sqrt{c_1c_2}}{1-e^{-\theta\epsilon/2}} < \infty.
\end{aligned}$$

This shows that flow B is also λ-m.b.f.a.

Appendix D
Proof of Proposition 12

To prove Proposition 12, we introduce the concept of work conserving modes in [9] for queues that have *at most one packet departure* in a time slot.

Definition 17 *[9] **(WC(K, D) queues)** A queue is in the work conserving (WC) mode if there is one departure in each time slot whenever the queue is nonempty. A queue is work conserving with response workload K and response delay D (denoted by $WC(K, D)$) if it has the following property: when the queue length is smaller than K at time $t-1$ and becomes longer than or equal to K at time t, this queue begins to be in the WC mode not later than time $t + D$. Moreover, this mode must continue until the queue length becomes smaller than K again.*

Clearly, each output buffer of an output-buffered switch is in the work conserving mode at every time slot. It is shown in [9] that there is a bound between the queue length of a WC queue and that of a $WC(K, D)$ queue.

Lemma 18 *[9] Let $Q_{WC}(t)$ (resp., $Q_{WC(K,D)}(t)$) be the number of packets in a WC (resp., $WC(K, D)$) queue at time t. Suppose that both queues are subject to the same arrival process and they both are empty at time 0. Then*

$$Q_{WC(K,D)}(t) \leq Q_{WC}(t) + K + D - 1. \tag{30}$$

We have the following work conserving property for the input buffers of a switch in the first stage in Figure 1.

Lemma 19 *Each input buffer of a switch in the first stage in Figure 1 is work conserving with response workload $N(p-1)+1$ and response delay $p-1$.*

Proof. Note that if there are more than $N(p-1)$ packets in an input buffer of a switch in the first stage in Figure 1, then there is at least one full-framed VOQ in that input buffer. As such, that input buffer will send out p packets during the next frame and it will continue to do so until there are no full-framed VOQs in that input buffer, at which point of time there are at most $N(p-1)$ packets in that input buffer. Therefore, each input buffer of a switch in the first stage in Figure 1 is work conserving with response workload $N(p-1)+1$. As the time it takes to the beginning time slot of the next frame is at most $p-1$, we see that the response delay is $p-1$. ■

Proof. (Proof of Proposition 12) Note that there is at most one packet arrival at an input buffer of a switch in the first stage in Figure 1 at any time. If we put the same arrival process to a work conserving queue, then the number of packets in that work conserving queue is at most 1. Thus, we have from Lemma 18 and Lemma 19 that the total number of packets in an input buffer of a switch in the first stage in Figure 1 is bounded above by

$$1+(N(p-1)+1)+(p-1)-1=N(p-1)+p.$$

The proof is completed. ■

References

[1] H. Ahmadi and W. E. Denzel, "A survey of modern high-performance switching techniques," *IEEE Journal on Selected Areas in Communications*, vol. 7, pp. 1091–1103, September 1989.

[2] S. Iyer and N. McKeown, "Making parallel packet switch practical," in *Proceedings IEEE Annual Conference on Computer Communications (INFOCOM'01)*, Anchorage, AK, USA, April 22–26, 2001.

[3] I. Stoica and H. Zhang, "Exact emulation of an output queueing switch by a combined input output queueing switch," in *Proceedings IEEE/IFIP International Workshop on Quality of Service (IWQoS'98)*, Napa, CA, USA, May 18–20, 1998, pp. 218–224.

[4] S.-T. Chuang, A. Goel, N. McKeown, and B. Prabhkar, "Matching output queueing with a combined input/output-queued switch," *IEEE Journal on Selected Areas in Communications*, vol. 17, pp. 1030–1039, June 1999.

[5] C.-S. Chang, D.-S. Lee, and Y.-S. Jou, "Load balanced Birkhoff-von Neumann switches, part I: one-stage buffering," *Computer Communications*, vol. 25, pp. 611–622, April 2002.

[6] I. Keslassy, S.-T. Chuang, K. Yu, D. Miller, M. Horowitz, O. Solgaard, and N. McKeown, "Scaling internet routers using optics," in *Proceedings ACM Conference of the Special Interest Group on Data Communication (SIGCOMM'03)*, Karlsruhe, Germany, August 25–29, 2003.

[7] Y. Shen, S. Jiang, S. S. Panwar, and H. J. Chao, "Byte-focal: a practical load-balanced switch," in *Proceedings IEEE Workshop on High Performance Switching and Routing (HPSR'05)*, Hong Kong, China, May 12–14, 2005.

[8] J.-J. Jaramillo, F. Milan, and R. Srikant, "Padded frames: a novel algorithm for stable scheduling in load-balanced switches," *IEEE/ACM Transactions on Networking*, vol. 16, pp. 1212–1225, October 2008.

[9] C.-L. Yu, C.-S. Chang, and D.-S. Lee, "CR switch: A load-balanced switch with contention and reservation," *IEEE/ACM Transactions on Networking*, vol. 17, pp. 1659–1671, October 2009. Conference version appeared in *IEEE INFOCOM* 2007.

[10] C. Clos, "A study of nonblocking switching networks," *Bell System Technical Journal*, vol. 32, pp. 406–424, March 1953.

[11] V. E. Benes, *Mathematical Theory of Connecting Networks and Telephone Traffic*, New York, NY: Academic Press, 1965.

[12] C.-S. Chang, J. A. Thomas, and S.-H. Kiang, "On the stability of open networks: a unified approach by stochastic dominance," *Queueing Systems: Theory and Applications*, vol. 15, pp. 239–260, March 1994.

[13] F. P. Kelly, "Notes on effective bandwidths," in *Stochastic Networks: Theory and Applications*, F. P. Kelly, S. Zachary, and I. Ziedins, Eds. Oxford, UK: Oxford University Press, 1996, pp. 141–168.

[14] D. V. Lindley, "The theory of queues with a single server," *Mathematical Proceedings of the Cambridge Philosophical Society*, vol. 48, pp. 277–289, April 1952.

[15] R. M. Loynes, "The stability of a queue with non-independent interarrival and service times," *Mathematical Proceedings of the Cambridge Philosophical Society*, vol. 58, pp. 497–520, July 1962.

[16] C.-S. Chang, "Stability, queue length and delay of deterministic and stochastic queueing networks," *IEEE Transactions on Automatic Control*, vol. 39, pp. 913–931, May 1994.

[17] R. L. Cruz, "A calculus for network delay, part I: Network elements in isolation," *IEEE Transactions on Information Theory*, vol. 37, pp. 114–131, January 1991.

[18] N. McKeown, V. Anantharam, and J. Walrand, "Achieving 100% throughput in an input-queued switch," in *Proceedings IEEE Annual Conference on Computer Communications (INFOCOM'96)*, San Francisco, CA, USA, March 24–28, 1996, pp. 296–302.

[19] L. G. Valiant, "A scheme for fast parallel communication," *SIAM Journal on Computing*, vol. 11, pp. 350–361, May 1982.

[20] C.-S. Chang, D.-S. Lee, Y.-J. Shih, and C.-L. You, "Mailbox switch: a scalable two-stage switch architecture for conflict resolution of ordered packets," *IEEE Transactions on Communications*, vol. 56, pp. 136–149, January 2008.

[21] C.-S. Chang and D.-S. Lee, "Quasi-circuit switching and quasi-circuit switches," in *Proceedings of IEEE International Conference on Information Technology: Research and Education (ITRE'05)*, Hsinchu, Taiwan, R.O.C., June 27–30, 2005.

[22] N. Chrysos and M. Katevenis, "Scheduling in non-blocking buffered three-stage switching fabrics," in *Proceedings IEEE Annual Conference on Computer Communications (INFOCOM'06)*, Barcelona, Spain, April 23–29, 2006.

[23] J. Turner, "An optimal non-blocking multicast virtual circuit switch," in *Proceedings IEEE Annual Conference on Computer Communications (INFOCOM'94)*, Toronto, Ontario, Canada, June 12–16, 1994.

[24] C. E. Koksal, "Providing quality of service in high speed electronic and optical switches," Ph.D. Dissertation, Massachusetts Institute of Technology, Cambrige, MA, USA, September 2002.

[25] M. J. Neely, E. Modiano, and Y.-S. Cheng, "Logarithmic delay for $N \times N$ packet switches under the crossbar constraint," *IEEE Transactions on Networking*, vol. 15, pp. 657–668, June 2007.

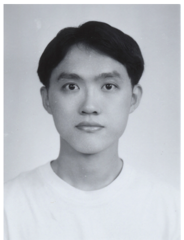
[26] R. B. Magill, C. E. Robhrs, and R. L. Stevenson, "Output-queued switch emulation by fabrics with limited memory," *IEEE Journal of Selected Areas in Communications*, vol. 21, pp. 606–615, May 2003.

[27] S. T. Chuang, S. Iyer, and N. McKeown, "Practical algorithms for performance guarantees in buffered crossbar switches," in *Proceedings IEEE Annual Conference on Computer Communications (INFOCOM'05)*, Miami, FL, USA, March 13–17, 2005.

[28] Q. Duan and J. Daigle, "Resource allocation for quality of service provision in multistage buffered crossbar switches," *Elsevier Journal of Computer Networks*, vol. 46, pp. 147–168, October 2004.

[29] Q. Duan, "Quality of service provision in combined input and crosspoint queued switches without output queuing match," *Elsevier Journal of Computer Communications*, vol. 30, pp. 830–840, February 2007.

Cheng-Shang Chang (S'85-M'86-M'89-SM'93-F'04) received the B.S. degree from National Taiwan University, Taipei, Taiwan, R.O.C., in 1983, and the M.S. and Ph.D. degrees from Columbia University, New York, NY, USA, in 1986 and 1989, respectively, all in Electrical Engineering. From 1989 to 1993, he was a Research Staff Member at the IBM Thomas J. Watson Research Center, Yorktown Heights, NY, USA. Since 1993, he has been with the Department of Electrical Engineering at National Tsing Hua University, Hsinchu, Taiwan, R.O.C., where he is currently a Professor. His current research interests are in high speed switching, communication network theory, and mathematical modeling of the Internet. Dr. Chang received an IBM Outstanding Innovation Award in 1992, an IBM Faculty Partnership Award in 2001, and Outstanding Research Awards from the National Science Council, Taiwan, R.O.C., in 1998, 2000 and 2002, respectively. He also received Outstanding Teaching Awards from both the college of EECS and the university itself in 2003. He was appointed as the first Y. Z. Hsu Scientific Chair Professor in 2002. He is the author of the book "Performance Guarantees in Communication Networks" and a coauthor of the book "Principles, Architectures and Mathematical Theory of High Performance Packet Switches." He served as an editor for Operations Research from 1992 to 1999. Dr. Chang is a member of IFIP Working Group 7.3 and is an editor for IEEE/ACM Transactions on Networking.



Jay Cheng (S'00-M'03-SM'09) received the B.S. and M.S. degrees from National Tsing Hua University, Hsinchu, Taiwan, R.O.C., in 1993 and 1995, respectively, and the Ph.D. degree from Cornell University, Ithaca, NY, USA, in 2003, all in Electrical Engineering. In August 2003, he joined the Department of Electrical Engineering at National Tsing Hua University, Hsinchu, Taiwan, R.O.C., where he is currently an Associate Professor. Since October 2004, he has also been affiliated with the Institute of Communications Engineering at National Tsing Hua University, Hsinchu, Taiwan, R.O.C. His current research interests include optical queueing theory, switching theory, communications theory, and information theory.

Duan-Shin Lee (S'89-M'90-SM'98) received the B.S. degree from National Tsing Hua University, Hsinchu, Taiwan, R.O.C., in 1983, and the M.S. and Ph.D. degrees from Columbia University, New York, NY, USA, in 1987 and 1990, respectively, all in Electrical Engineering. He was a research staff member at the C&C Research Laboratory of NEC USA, Inc., Princeton, NJ, USA, from 1990 to 1998. He joined the Department of Computer Science at National Tsing Hua University, Hsinchu, Taiwan, R.O.C., in 1998, where he is currently a Professor. His research interests are in high-speed switch and router design, wireless networks, performance analysis of communication networks, and queueing theory.

Chi-Feung Wu received the B.S. degree in Electrical Engineering and the M.S. degree in Communications Engineering from National Tsing Hua University, Hsinchu, Taiwan, R.O.C., in 2006 and 2008, respectively. In August 2008, he served as an international volunteer for information and communication technology training in Ghana in Africa. He has been serving his military service since October 2008.