Available at www.ComputerScienceWeb.com
POWERED BY SCIENCE DIRECT®

ELSEVIER

Performance Evaluation 55 (2004) 113–128

PERFORMANCE EVALUATION
An International Journal

www.elsevier.com/locate/peva

Delay bounds for combined input–output switches with low speedup ☆

Paolo Giaccone [a,*], Emilio Leonardi [a], Balaji Prabhakar [b], Devavrat Shah [b]

[a] *EE Department, Politecnico di Torino, Torino, Italy*
[b] *EE/CS Departments, Stanford University, Stanford, CA 94305, Italy*

Abstract

The *speedup* of a switch is the factor by which the switch, and hence the memory used in the switch, runs faster compared to the line rate. In high-speed switches, line rates are already touching limits at which memory can operate. In this scenario, it is very important for a switch to run at as low a speedup as possible.

In the past, it has been shown that 100% throughput can be achieved for any admissible traffic for an input queued (IQ) switch [IEEE Trans. Commun. 47 (8) (1999) 1260; The throughput of data switches with and without speedup, in: Proceedings of the IEEE INFOCOM'00, vol. 2, Tel Aviv, Israel, March 2000, pp. 556–564] at speedup 1. This gives finite average delays but does not guarantee control on packet delays. In [IEEE J. Sel. Areas Commun. 17 (6) (1999) 1030], authors show that a combined input–output queued (CIOQ) switch can emulate perfectly an output queued (OQ) switch at a speedup of 2 and, thus, control the packet delays. This motivates the study of possibility of obtaining delay control at speedup less than 2. To guarantee optimal control of delays for a general class of traffic, as shown in [3], speedup 2 is necessary. Hence, to obtain control of delays at lower speedup, we need to restrict the class of arrival traffics. In this paper, we study the speedup requirement for a class of admissible traffic, which we will denote as $(1, nF)$-regulated traffic, with parameters n and F. We obtain the necessary speedup for this class of traffic. Further, we present a general class of algorithms working at the necessary speedups and providing bounded delays.

Keywords: Combined input–output switches; Scheduling; Delay bounds; Speedup

1. Introduction

Recently, input queued (IQ) and combined input–output queued (CIOQ) switches with virtual output queueing (VOQ) have become an attractive architectural solution in very high speed routers [4,5] as they scale well with the line rate.

At the same time, output queued (OQ) switches are attractive as they achieve 100% throughput under any admissible traffic and give control over delays. But OQ switches require memory bandwidth (at the

☆ A preliminary version of this paper has been published in the Proceedings of IEEE Globecom 2002.

* Corresponding author. Tel.: +39-011-5644036; fax+39-011-5644099.
E-mail addresses: giaccone@polito.it (P. Giaccone), leonardi@polito.it (E. Leonardi), balaji@isl.stanford.edu (B. Prabhakar), devavrat@cs.stanford.edu (D. Shah).

0166-5316/$ – see front matter © 2003 Elsevier B.V. All rights reserved.
doi:10.1016/S0166-5316(03)00103-2

output ports) to scale as O(rN), where r is the line rate and N is the number of ports. In other words, the internal switching speed has to run N times faster than the line rate, that is, speedup S is N. This constrains the speed at which OQ switches can run.

A pure IQ switch is able to achieve very high speeds, since the memory bandwidth scales as O(r), being by construction its speedup equal to 1. The main drawback of this architecture is that it requires a scheduling algorithm which selects a non-conflicting set of packets to transfer across the switch. This scheduling algorithm should be simple, because it is implemented in hardware at very high speed. A class of maximum weight matching (MWM) algorithms for IQ switches are known which provide 100% throughput for any admissible traffic [1,2,6]. In [7,8] bounds on the average delay are obtained for MWM algorithm under admissible Bernoulli i.i.d. traffic pattern. But they do not guarantee delay bounds for each packet. Many practical scheduling algorithms [9,10] have been proposed to approximate MWM performance. Their simplicity usually leads to some performance penalties, usually in the form of throughput degradation and/or larger delays.

In [2,11] it is shown that at speedup 2, simple maximal matching kind of algorithms are stable (provide 100% throughput) under admissible arrival traffic. But again, there are no strict delay guarantees provided. In [3] it is shown that $S \geq 2$ is necessary and sufficient to emulate performance of OQ switches and, thus, to control the delays. Unfortunately the perfect emulation of OQ requires complicated stable-marriage style algorithms which are not feasible to implement at a very high-speed. In [12] it was shown that simpler scheduling algorithms can achieve the same performance of an OQ switch in terms of average delay.

Since speedup higher than 1 limits the speed at which a switch can operate, it is very desirable to operate at as low speedup as possible. This leads us to investigate a possible *tradeoff between speedup and delay*. However, if we want to obtain delay control for speedup $1 \leq S < 2$, we must restrict the arrival traffic. In this paper, we consider a general enough class of arrival traffic and study the necessary and sufficient speedup $1 \leq S < 2$ required to emulate OQ performance with guaranteed delay bounds.

The rest of the paper is organized as follows. In Section 2.1, we define the architecture of the CIOQ switch which is of our interest. In Section 2.2 we present some important definitions. Section 2.3 deals with notations used in the later part of the paper. Section 2.4 defines the restricted traffic class we consider in this paper. In Section 3 we consider the essential properties of an F-work-conserving switch, which incurs at most a delay penalty of F compared to OQ switch. Section 4 talks about the necessary and sufficient speedup of a CIOQ switch to emulate F-work-conservation.

2. Basic model, definitions and notations

2.1. A CIOQ switch

An $N \times N$ CIOQ switch has N inputs and N outputs with cross-bar in the switch fabric, as shown in Fig. 1. The queues at each input are logically divided into N virtual output queues (VOQ) corresponding to N different outputs. There are queues at outputs too. When a CIOQ switch is working at speedup S (with $1 \leq S \leq N$), each input is able to transfer up to S packets per time slot, and each output is able to receive up to S packets per time slot. At speedup $S = 1$ a CIOQ switch is same as IQ switch, and does not require queues at the output side.

We assume that time is slotted. In a given time slot, at most one packet can arrive at each input. In every "scheduling cycle", the cross-bar can transfer one packet from each input and one packet to each

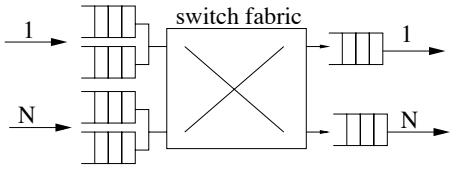

Fig. 1. Architecture of an $N \times N$ CIOQ.

output. Effectively for a CIOQ switch operating at a speedup S, S scheduling cycles happen during one time slot. For example, if $S = 3/2$, then every one time slot 1.5 scheduling cycles happen. That is, in real switch every two time slots, three scheduling cycles happen.

2.2. *Work conservation*

Next we would like to consider the concept of work conservation for a switch. Consider the following definition, which was first proposed in [12] motivated from the classical queueing theory.

Definition 1. A switch is work-conserving if and only if, for any time slot, an output is always transferring one packet to the outgoing link whenever a packet is present in the system directed to the considered output.

Note that this definition requires that the system should be "observed" at each time slot to check if it is work-conserving.

An OQ switch is by construction work-conserving whereas an IQ switch is not work-conserving. For example, consider a 3×3 IQ switch in which at time $t = 0$ no backlog exists and at time $t = 1$ two packets arrive: one at input 1 directed to output 3 and one at input 2 directed to output 3. An arrived packet is immediately transferred to the outputs and transmitted, while the other packet is stored at the input. At time $t = 2$ other two packets arrive: one packet at input 1 directed to output 2 and one packet at input 2 directed to output 1. Now at the inputs there are three packets directed to different outputs, but only two of them can be transferred to the outputs thus an output port remains idle even if there is a packet directed to it. As a conclusion an IQ switch may be non work-conserving. Note that a work-conserving switch ensures the minimum average delays, (i.e. the same average delay than an OQ switch) since an output is never idling as long as a packet directed to it is in the switch.

The work-conserving property of OQ switch suggests the following equivalent work conservation property which was first considered in [3].

Definition 2. A switch, in particular CIOQ switch, is work-conserving iff, for any arrival sequence $\mathcal{A}$ the following holds for all the time slots: for each output j, the number of packets in the switch waiting for transmission to j equals the number of packets that would be stored in an OQ under the same $\mathcal{A}$.

From [3], speedup 2 is necessary to emulate OQ and hence to be strictly work-conserving for a CIOQ switch. The goal of this paper is to consider the switch operating at speedup $1 \leq S < 2$ while providing bounds on performance difference between CIOQ switch and an OQ switch. This leads to the notion of little less strict work-conserving property which we call as F-work conservation. Basically, instead

of requiring the system to be work-conserving every time, we consider system with property of work conservation holding at every F times.

Definition 3. A CIOQ switch is F-work-conserving iff, for any arrival sequence $\mathcal{A}$ the following holds for time $t = 0, F, 2F, \ldots, kF, \ldots$: for each output j the number of packets in the switch waiting for transmission directed to output j equals the number of packets that would be stored in an OQ under the same $\mathcal{A}$. We call the time interval $\{t \in \mathbb{Z}^+ : t \in [(k-1)F + 1, kF]\}$ as the kth observation window.

The most important property about F-work-conserving switches is about the control of the delays. We compare the delays experienced by packets in a CIOQ switch with an F-work-conserving policy and in an OQ switch under the same arrival sequence.

Theorem 1. *Fix any admissible arrival traffic sequence $\mathcal{A}$ at a switch of size N. Suppose an OQ switch and an F-work-conserving CIOQ switch are given the same arrival traffic pattern $\mathcal{A}$. For any packet $P \in \mathcal{A}$, let T_{OQ}^P be the departure time from the OQ switch. Similarly, let $T_{\mathcal{D}}^P$ be the departure time of the same packet P under the F-work-conserving CIOQ switch. Then for every P departing from OQ switch, there exists a unique packet $P' \in \mathcal{A}$ departing from CIOQ switch from the same output as P, such that*

$$T_{\mathcal{D}}^{P'} - T_{OQ}^{P} \leq F - 1. \tag{1}$$

Hence, the average delay per packet experienced by F-work-conserving CIOQ switch is at most $F - 1$ times more than the OQ switch for each feasible traffic pattern $\mathcal{A}$.

Proof. We apply exactly the same traffic sequence $\mathcal{A}$ to both: (a) an OQ switch, and (b) an F-work-conserving CIOQ switch.

We would like to prove the statement by induction. At time $t = 0$, both systems start empty and hence the statement is trivially true. Assume that the theorem statement is true for all packets departing from OQ till time kF. By F-work conservation property, the number of packets queued for any of the output in both OQ and CIOQ switch is the same at time kF. Consider $P_1, \ldots, P_m$ packets departed from output j in OQ switch between time $kF + 1, \ldots, (k + 1)F$, where $m \leq F$, depending on arrival pattern $\mathcal{A}$. Since

- at the end of time kF, both OQ and CIOQ had the same number of packets enqueued for output j,
- at the end of time $(k + 1)F$, both OQ and CIOQ have the same number of packets enqueued for output j, and
- there are m packets $P_1, \ldots, P_m$ departing from output j in OQ switch between time $kF+1, \ldots, (k+1)F$,
- there are m packets $P'_1, \ldots, P'_m$ departing from output j of CIOQ by the end of time $(k + 1)F$.

We can associate each of the P_i with unique P'_i and obtain,

$$T_{\mathcal{D}}^{P'_i} - T_{OQ}^{P_i} \leq F - 1,$$

which means that the average departure time in CIOQ differs at most by $F - 1$ from OQ. Then the same property holds for the average delay, since the arrival sequence is the same for CIOQ and OQ. This completes the proof of Theorem 1. □

We would like to note that Theorem 1 refers to a much stronger property than just a bounded average delays. For example, under admissible traffic an IQ switch running at speedup 1 and using MWM

scheduling policy has a bounded average delay, and hence bounded average delay with respect to OQ switch too (by definition OQ has average delay ≥ 0). But it does not imply the property of Theorem 1.

2.3. Notations

Consider an $N \times N$ CIOQ switch. We observe the system at times $t_k = kF$, $\forall k \in \mathbb{Z}^+$, since we are interested in F-work-conserving property. We define the following notations:

- B_{ij}^k is the number of packets enqueued at the input port i and destined to output j, sampled at the beginning of the observation window k, at time $t = kF\ \forall k \in \mathbb{Z}^+$.
- $\hat{B}_j^k \triangleq \sum_i B_{ij}^k$ and $\bar{B}_i^k \triangleq \sum_j B_{ij}^k$.
- $A_{ij}(t)$ is the number of arrivals from input i to output j at time $t\ \forall t \in \mathbb{Z}^+$; $A(t) = [A_{ij}(t)]$. A_{ij}^k is the cumulative number of arrivals from input i to output j occurring during the $(k-1)$th observation window: $A_{ij}^k = \sum_{t=(k-1)F}^{kF-1} A_{ij}(t)$. $A^k = [A_{ij}^k]$.
- $\hat{A}_j^k \triangleq \sum_i A_{ij}^k$ and $\bar{A}_i^k \triangleq \sum_j A_{ij}^k$.
- D_{ij}^k is the cumulative number of services from input i to output j, occurring during the kth observation window: $D^k = [D_{ij}^k]$.
- $\hat{D}_j^k \triangleq \sum_i D_{ij}^k$ and $\bar{D}_i^k \triangleq \sum_j D_{ij}^k$.
- O_j^k is the number of packets enqueued at the output port j, sampled at the beginning of the kth observation window.
- $Y_j^k = \sum_i B_{ij}^k + O_j^k$ is the total number of packet queued in the system and destined to output j.
- $\lceil x \rceil^+ = \max\{0, x\}$.

To model the system, we consider the switch evolving in a *gated-fashion* with period F, i.e. new arrivals are aggregated during each observation window and they are scheduled only at the beginning of the next observation window. It is like considering batch arrivals at the beginning of a new observation window, by batching all the arrivals during the previous observation window. The evolution of the state of the system is sampled at the beginning of a new observation window and can be modeled as follows:

$$B_{ij}^{k+1} = B_{ij}^k + A_{ij}^k - D_{ij}^k \quad \forall i, j, \tag{2}$$

$$O_j^{k+1} = \left\lceil O_j^k + \sum_i D_{ij}^k - F \right\rceil^+ \quad \forall j, \tag{3}$$

$$Y_j^{k+1} = \left\lceil Y_j^k + \hat{A}_j^k - F \right\rceil^+ . \tag{4}$$

Eq. (2) models the system evolving in a gated fashion. Indeed, the new backlogged packets are given by the old ones, plus the new arrivals and minus the departures, both occurring during the previous observation window. Note that, when $F = 1$, Eq. (2) degenerates into the evolution of a generic discrete-time queue. It is important to highlight that a system evolving in a gated-fashion can increase the delay of a packet by at most F time slots, with respect to a slot-by-slot system. Eqs. (3) and (4) describe the transfer of all the scheduled packets directed to a generic output; in fact, during each observation window, at most F packets can be transferred to the output line cards. Eq. (4) assumes implicitly that the considered CIOQ switch is F-work-conserving.

Define the following norm.

Definition 4 (IO norm). Given $X \in \mathbb{R}^{N^2}$:

$$\|X\|_{\mathrm{IO}} \triangleq \max\left\{\max_j\left\{\sum_i X_{ij}\right\}, \max_i\left\{\sum_j X_{ij}\right\}\right\}.$$

A policy $\mathcal{D}$ working with a speedup S is feasible if:

$$\|D^k\|_{\mathrm{IO}} \leq SF \quad \forall k, B_{ij}^k, A_{ij}^k. \tag{5}$$

Indeed, by Birkhoff von Neumann theorem, any set D^k can be scheduled [13] in a time window of $\|D^k\|_{\mathrm{IO}}$ slots, since D^k can be decomposed in $\|D^k\|_{\mathrm{IO}}$ switching configurations.

2.4. Traffic class

In our context, we consider only controlled traffic, since it is the only one for which it is possible to guarantee delay bounds in an OQ switch architecture. We consider here only two kinds of controlled traffic: regulated and leaky bucket constrained traffic. Since at most one packet arrives per time slot, the following property holds when the arrivals are observed at the inputs:

$$\bar{A}_i \leq F. \tag{6}$$

2.4.1. Regulated traffic

The following definition is derived by the adversary queueing theory [14].

Definition 5. An arrival process $\mathcal{A}$ is (ρ, W)-regulated if:

$$\left\|\sum_{z=t}^{t+W-1} A(z)\right\|_{\mathrm{IO}} \leq \rho W \quad \forall t,$$

i.e., at most ρW packets arrive during each interval of W time slots for each input–output couple. W is called "admissibility window".

We can say that a (ρ, W)-regulated traffic injects at most ρW packets during an admissibility window W, corresponding to a maximum average rate ρ for each input–output couple during the same window W. Furthermore, an arrival process (ρ, W)-regulated is also $(1, \rho W)$-regulated, but not vice versa. In other words, the family of all the possible arrival processes (ρ, W)-regulated is a subset of the bigger family of processes $(1, \rho W)$-regulated.

We focus on $(1, nF)$-regulated arrival processes for which it holds:

$$\left\|\sum_{z=k}^{k+n-1} A^z\right\|_{\mathrm{IO}} \leq nF \quad \forall k. \tag{7}$$

2.4.2. Leaky bucket constrained traffic

This second kind of source is the leaky bucket constrained [15].

Definition 6. An arrival process $\mathcal{A}$ is $[\rho, \sigma]$-LBC (leaky bucket constrained) with leaky rate $\rho < 1$ and bucket size σ if

$$\left\| \sum_{t=k}^{k+n-1} A(t) \right\|_{\mathrm{IO}} \leq \gamma_{\rho,\sigma}(n) \quad \forall k,$$

where, being $u(t)$ the step function,

$$\gamma_{\rho,\sigma}(t) = (\sigma + \rho t)u(t).$$

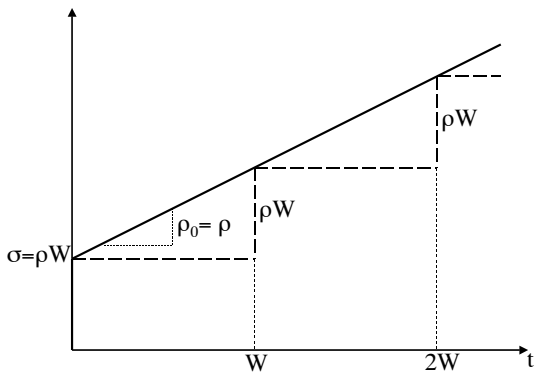

Fig. 2. Any (ρ, W)-regulated arrival process (the dashed stair-case is the worst case) can be seen as a $[\rho_0, \sigma]$-LBC arrival process (the solid line is the limit for such traffic).

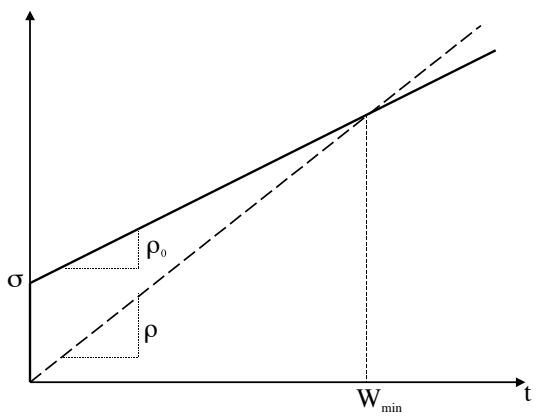

Fig. 3. Any $[\rho_0, \sigma]$-LBC arrival process (solid line) can be seen as a particular (ρ, W)-regulated traffic pattern (dashed line is the function ρt) when $W \geq W_{\min}$.

Note that a $[\rho, \sigma]$-LBC source is also $(1, nF)$-regulated, with the following choice of F:

$$F \geq \left\lceil \frac{\sigma}{n(1-\rho)} \right\rceil, \tag{8}$$

where $\lceil x \rceil$ is the smallest integer greater or equal than x and (8) is given by solving $\gamma_{\rho,\sigma}(nF) \leq nF$: under these conditions, we have that the arrival curve for a regulated source is always above the arrival curve corresponding to the LBC source. Figs. 2 and 3 show the relations between (ρ, W)-regulated and $[\rho_0, \sigma]$-LBC arrival processes.

3. Properties of F-work-conserving policies

Property 1. *A policy $\mathcal{D}$ is F-work-conserving in an observation window of size F with speedup S if*

$$\hat{B}_j^{k+1} \leq \left\lceil \hat{B}_j^k + \hat{A}_j^k + O_j^k - F \right\rceil^+ \quad \forall k, j. \tag{9}$$

To understand the meaning of this property, start to consider the case $F = 1$. Eq. (9) means that if at least a packet is present at the input ports destined for output j, this (single) packet should be transferred to the output queue j, provided that no packet at the output queue j is present. For a generic F, Eq. (9) implies that, if at least $F - O_j^k$ packets are present at the input ports destined for output j, these packets should be transferred to the output queue j.

For F-work-conserving policies we state the following theorem.

Theorem 2. *Assume that policy $\mathcal{D}$ is F-work-conserving and the arrival process $\mathcal{A}$ is $(1, nF)$-regulated. If $Y_j^k > 0$ then*

$$\exists n_0 : 0 \leq n_0 < n, \qquad Y_j^k Y_j^{k+n_0} = 0,$$

i.e., it exists a k' close to k (that is, $k' - k < n$) such that $Y_j^{k'} = 0$.

Proof. *Case $n = 1$.* Here, the meaning of the theorem is that $Y_j^k = 0$ for all k. By induction, assume that the theorem holds for an epoch less or equal to k, hence $Y_j^k = 0$. Recall (4) and, by contradiction, assume $Y_j^{k+1} > 0$:

$$Y_j^k = 0, \qquad 0 < Y_j^{k+1} \leq \left\lceil Y_j^k + \hat{A}_j^k - F \right\rceil^+ \leq \hat{A}_j^k - F \Rightarrow \hat{A}_j^k > F,$$

which is in contradiction with $\mathcal{A}$ which is $(1, F)$-regulated.

Case $n = 2$. By induction, assume that the theorem holds for an epoch less or equal to $k - 1$, hence $Y_j^{k-1} = 0$ and $Y_j^k > 0$. Recall (4) and by contradiction, assume $Y_j^{k+1} > 0$:

$$0 < Y_j^k \leq \left\lceil Y_j^{k-1} + \hat{A}_j^{k-1} - F \right\rceil^+ = \hat{A}_j^{k-1} - F,$$

$$0 < Y_j^{k+1} \leq \left\lceil Y_j^k + \hat{A}_j^k - F \right\rceil^+ \leq \hat{A}_j^{k-1} - F + \hat{A}_j^k - F \Rightarrow \hat{A}_j^k + \hat{A}_j^{k-1} > 2F.$$

But, since $\mathcal{A}$ is $(1, 2F)$-regulated, $\hat{A}_j^k + \hat{A}_j^{k-1} \leq 2F$, which is in contradiction. The proof can be easily extended for $n > 2$. □

Note that Theorem 2 implies that the maximum delay experienced by packets of an $(1, nF)$-regulated arrival process in a CIOQ switch with an F-work-conserving policy is not greater than nF slots.

We now show one possible example of F-work-conserving policy.

Lemma 1. *The following policy* $\mathcal{D}$:

$$D_{ij}^k = (A_{ij}^k + B_{ij}^k) \min\left\{1, \frac{\theta F - \gamma O_j}{\hat{A}_j^k + \hat{B}_j^k}\right\} \quad \forall i, j, k$$

is F-work-conserving for $\theta \geq 1$ *and* $0 \leq \gamma \leq 1$.

Proof. If $\hat{A}_j^k + \hat{B}_j^k \leq \theta F - \gamma O_j$, then $\hat{B}_j^{k+1} = 0$ and $\hat{D}_j^k = \hat{B}_j^k + \hat{A}_j^k$. Otherwise, if $\hat{A}_j^k + \hat{B}_j^k > \theta F - \gamma O_j$, then $\hat{B}_j^{k+1} = \hat{B}_j^k + \hat{A}_j^k - \theta F + \gamma O_j > 0$ and $\hat{D}_j^k = \theta F - \gamma O_j$. Hence, if $\theta \geq 1$ and $\gamma \in [0, 1]$,

$$\hat{B}_j^{k+1} \leq \left\lceil \hat{B}_j^k + \hat{A}_j^k - \theta F + \gamma O_j \right\rceil^+ \leq \left\lceil \hat{B}_j^k + \hat{A}_j^k - F + \gamma O_j \right\rceil^+ \leq \left\lceil \hat{B}_j^k + \hat{A}_j^k - F + O_j \right\rceil^+,$$

and the policy $\mathcal{D}$ is F-work-conserving. □

Policy $\mathcal{D}$, to be feasible with the speedup S, satisfies the following relation, derived from Eq. (5), referred as *feasibility condition*:

$$SF \geq \|D^k(\theta, \gamma)\|_{\mathrm{IO}} \quad \forall k.$$

Intuitively, policy $\mathcal{D}$, with $\gamma = 0$, is greedy, since it transfers completely all the backlogged packets if compatible with the available output bandwidth θF. Otherwise, the output bandwidth is distributed among all the inputs proportionally to the number of backlogged packets.

4. On the minimum speedup under regulated traffic

The following three theorems are our main results. The first one is quite trivial and intuitive, but can be significant.

Theorem 3. *Consider a CIOQ switch. Under an arrival process* $\mathcal{A}$ *which is* $(1, W)$*-regulated, there exists a W-work-conserving policy when* $S \geq 1$.

Proof. Fix the observation window size $F = W$. Consider the following policy:

$$D_{ij}^k = (A_{ij}^k + B_{ij}^k) \min\left\{1, \frac{F}{\hat{A}_j^k + \hat{B}_j^k}\right\}.$$

We know, from Lemma 1, that it is F-work-conserving (in this case, $\theta = 1$ and $\gamma = 0$). Now we will prove that it is feasible for $S \geq 1$. Thanks to Theorem 2, we can assume, for all k,

$$Y_j^k = 0 \quad \Rightarrow \quad B_{ij}^k = 0 \quad \forall i \text{ and } O_j^k = 0.$$

By assumption, $\hat{A}_j^k \leq F$ and $\bar{A}_i^k \leq F$. Hence, the policy reduces to: $D_{ij}^k = A_{ij}^k$ and by imposing $\|D^k\|_{\text{IO}} \leq SF$, we obtain $S \geq 1$. □

Theorem 4. *Consider a CIOQ switch. Under an arrival process $\mathcal{A}$ which is $(1, W)$-regulated, there exists a $W/2$-work-conserving policy if and only if $S \geq 4/3$.*

Proof. Fix the observation window size $F = W/2$. We divide the proof in two steps, in the first we show that $S = 4/3$ is a sufficient speedup to deal with $(1, 2F)$-regulated traffic, in the second step we show that it is also a necessary condition. Note that in this case, $\mathcal{D}$ is also the optimal policy, minimizing the speedup needed.

Step 1. Fix $\theta_0 = 4/3$ and consider the following policy $\mathcal{D}$:

$$D_{ij}^k = (A_{ij}^k + B_{ij}^k) \min\left\{1, \frac{\theta_0 F}{\hat{A}_j^k + \hat{B}_j^k}\right\}.$$

We know, from Lemma 1, that $\mathcal{D}$ is F-work-conserving (in this case, $\gamma = 0$ and $\theta = \theta_0$), hence it is a good representative for $\mathcal{D}$. We show now that $\mathcal{D}$ is feasible for $S \geq 4/3$. First we notice that, in general,

$$\hat{D}_j^k = \sum_i D_{ij}^k = \min\{\hat{A}_j^k + \hat{B}_j^k, \theta_0 F\} \leq \theta_0 F \leq SF$$

with $S \geq 4/3$. Thus, to decide the feasibility of $\mathcal{D}$, we have to compute the maximum possible value for $\bar{D}_i^k$. $\bar{D}_i^k$ can be splitted in two components, $\bar{D}_{i,A}^k$ which is the amount of services received by packets arrived during the kth observation window at input i, and $\bar{D}_{i,B}^k$ is the amount of services received by backlogged packets from the previous observation window at input i: $\bar{D}_i^k = \bar{D}_{i,A}^k + \bar{D}_{i,B}^k$. It is $\bar{D}_{i,A}^k \leq F$ because of (6). We now find the maximum for $\bar{D}_{i,B}^k$. Note that if $\hat{D}_{j,B}^k > 0$ then $\hat{B}_j^k > 0$, being $\hat{D}_{j,B}^k$ the amount of service received by backlogged packets at output j. Then, $\hat{B}_j^{k-1} = 0$ and $D_{ij,B}^k = B_{ij}^k$ for Theorem 2.

$$\sum_j B_{ij}^k = \sum_j A_{ij}^{k-1}\left(1 - \min\left\{1, \frac{\theta_0 F}{A_j^{k-1}}\right\}\right) \leq \sum_j A_{ij}^{k-1}\left(1 - \min\left\{1, \frac{\theta_0 F}{2F}\right\}\right) \leq F\left(1 - \frac{\theta_0}{2}\right)$$

thanks to the fact that $A_j^{k-1} \leq 2F$. Thus, after maximizing $\bar{D}_{i,B}^k$, we can maximize $\bar{D}_i^k$ and imposing the feasibility conditions:

$$\bar{D}_i^k \leq F + F(1 - \tfrac{1}{2}\theta_0) = \tfrac{4}{3}F \leq SF,$$

which holds for $S \geq 4/3$.

In conclusion, with speedup $S \geq 4/3$ policy $\mathcal{D}$ is feasible.

Step 2. We want to show, by a counterexample, that the minimum speedup 4/3 is also necessary to have an F-work-conserving policy. Consider a switch with two active inputs and three outputs. Assume $Y_j^k = 0$, hence $B_{ij}^k = 0$ for $1 \leq i \leq 2$ and $1 \leq j \leq 3$. Consider the following traffic pattern, $(1, 2F)$-regulated: $A_{11}^k = A_{21}^k = A_{12}^{k+1} = A_{23}^{k+1} = F$. At the end of the kth observation window, to minimize the maximum backlog at both inputs, we set $D_{11}^k = D_{21}^k = SF/2$.

After the arrival at time $k+1$, there are $(1-S/2)F$ packets enqueued at the inputs and destined to output 1, whereas F are destined to output 2 and 3. Hence, to have $\mathcal{D}$ work-conserving by setting $Y_j = 0$ and $B_{ij}^{k+2} = 0$: $D_{ij}^{k+1} = B_{ij}^{k+1} + A_{ij}^{k+1}$. Since D_{ij}^{k+1} must be feasible, we impose

$$\frac{(2-S)F}{N} + \frac{2(2N-1)F}{3N} \leq SF \quad \Rightarrow \quad S \geq \frac{4}{3}.$$

Hence, $S \leq 4/3$ is a necessary condition to have an F-work-conserving policy. □

Theorem 5. *Consider a CIOQ switch. Under an arrival process $\mathcal{A}$ which is $(1, W)$-regulated, there exists a $W/3$-work-conserving policy, if $S \geq 3/2$.*

Proof. Fix the observation window size $F = W/3$. Consider the following policy $\mathcal{D}$:

$$D_{ij}^k = (A_{ij}^k + B_{ij}^k) \min\left\{1, \frac{\theta F}{\hat{A}_j^k + \hat{B}_j^k}\right\}.$$

Note that, thanks to Lemma 1, $\mathcal{D}$ is F-work-conserving (in this case, $\gamma = 0$). We show now that $S \geq 3/2$ is a sufficient and necessary condition for $\mathcal{D}$ to be feasible. In this case, the considered policy may not be the optimal policy, minimizing the speedup needed.

Step 1. Fix $S \geq 3/2$. We first observe the following property.

Property 2. *It can never happen that it exists j such that $B_j^k > 0$ and $B_j^{k+1} > 0$.*

By contradiction, we can write the system evolution:

$$B_j^k = A_j^{k-1} - D_j^{k-1} + B_j^{k-1} > 0, \tag{10}$$

$$B_j^{k+1} = A_j^k - D_j^k + A_j^{k-1} - D_j^{k-1} > 0 \tag{11}$$

since $B_j^{k-1} = 0$ for Theorem 2. From both (10) and (11) we deduce that $D_j^k = D_j^{k+1} = SF$. But, because of the traffic features, it should be satisfied: $D_j^k + D_j^{k+1} = 2SF < A_j^k + A_j^{k+1} \leq 3F$, hence $S < 3/2$ which is in contradiction with our assumptions.

Consider the policy $\mathcal{D}$ and fix $\theta_0 = 3/2$. We know, from Lemma 1, that $\mathcal{D}$ is F-work-conserving (in this case, $\gamma = 0$ and $\theta = \theta_0$). We show now that $\mathcal{D}$ is feasible for $S \geq 3/2$. First we notice that, in general,

$$\hat{D}_j^k = \sum_i D_{ij}^k = \min\{\hat{A}_j^k + \hat{B}_j^k, \theta_0 F\} \leq \theta_0 F \leq SF$$

with $S \geq 3/2$. Thus, to decide the feasibility of $\mathcal{D}$, we have to compute the maximum possible value for $\bar{D}_i^k$. $\bar{D}_i^k$ can be splitted in only two components, thanks to Property 2, $\bar{D}_{i,A}^k$ which is the amount of services received by packets arrived during the observation window k at input i, and $\bar{D}_{i,B}^k$ is the amount of services received by backlogged packets from the previous observation window at input i: $\bar{D}_i^k = \bar{D}_{i,A}^k + \bar{D}_{i,B}^k$. It is $\bar{D}_{i,A}^k \leq F$ because of (6). We now find the maximum for $\bar{D}_{i,B}^k$, being $\hat{D}_{j,B}^k$ the amount of service received by backlogged packets at output j. Note that if $\hat{D}_{j,B}^k > 0$ then $\hat{B}_j^k > 0$. Then, $\hat{B}_j^{k-1} = 0$ and $D_{ij,B}^k = B_{ij}^k$ for Theorem 2.

$$\sum_j B_{ij}^k = \sum_j A_{ij}^{k-1} - D_{ij}^{k-1} = \sum_j A_{ij}^{k-1} \left(1 - \min\left\{1, \frac{\theta_0 F}{A_j^{k-1}}\right\}\right)$$
$$\leq \sum_j A_{ij}^{k-1} \left(1 - \min\left\{1, \frac{\theta_0 F}{3F}\right\}\right) \leq F\left(1 - \frac{\theta_0}{3}\right) \tag{12}$$

thanks to the fact that $A_j^{k-1} \leq 3F$. Thus, after maximizing $\bar{D}_{i,B}^k$, we can maximize $\bar{D}_i^k$ and imposing the feasibility conditions:

$$\bar{D}_i^k \leq F + F(1 - \tfrac{1}{3}\theta_0) = \tfrac{3}{2}F \leq SF,$$

which holds for $S \geq 3/2$.

In conclusion, with speedup $S \geq 3/2$ policy $\mathcal{D}$ is feasible.

Step 2. We now prove, by a counterexample, that the minimum speedup 3/2 is a necessary condition for $\mathcal{D}$ to be F-work-conserving. By contradiction, consider a switch with three active inputs and three active outputs, and j such that $Y_j^k = 0$: $B_{ij}^k = 0$ for $1 \leq i \leq 3$. We assume $S < 3/2$ that implies $\theta < S < 3/2$. Consider the following traffic pattern, $(1, 3F)$-regulated: $A_{11}^k = A_{21}^k = A_{31}^k = A_{12}^{k+1} = A_{22}^{k+1} = A_{32}^{k+1} = A_{13}^{k+2} = F$. At the end of observation window k, the service is given by: $D_{11}^k = D_{21}^k = D_{31}^k = \theta F/3$.

After the arrival at time $k+1$, there are $(1-\theta/3)F$ packets enqueued at each active input and destined to output 1, whereas F packets are destined to output 2. Hence, at time $k+1$ the policy $\mathcal{D}$ gives

$$D_{11}^{k+1} = D_{21}^{k+1} = D_{31}^{k+1} = \tfrac{1}{3}\theta F, \qquad D_{12}^{k+1} = D_{22}^{k+1} = D_{32}^{k+1} = \tfrac{1}{3}\theta F.$$

After the arrival at time $k+2$, there are $(1-2\theta/3)F$ packets enqueued at each active input and destined to output 1, $(1-\theta/3)F$ packets at each active input and destined to output 2, whereas F packets are destined to output 3. Hence, at time $k+2$ the policy $\mathcal{D}$ gives

$$D_{11}^{k+2} = D_{21}^{k+2} = D_{31}^{k+2} = (1 - \tfrac{2}{3}\theta)F, \qquad D_{12}^{k+2} = D_{22}^{k+2} = D_{32}^{k+2} = \tfrac{1}{3}\theta F, \qquad D_{13}^{k+2} = F.$$

To impose the feasibility of D^{k+2}, it should be:

$$D_{11}^{k+2} + D_{12}^{k+2} + D_{13}^{k+2} \leq SF \quad \Rightarrow \quad (2 - \tfrac{1}{3}\theta)F < \tfrac{3}{2}F$$

thus $\theta > 3/2$ which is in contradiction with our assumptions. Hence, $S \geq 3/2$ is a necessary condition for $\mathcal{D}$ to be feasible. □

It is possible to compute the minimum speedup necessary and sufficient for $\mathcal{D}$, with $\gamma = 0$, to be F-work-conserving under a generic $(1, nF)$-regulated traffic. The idea is to generalize the counterexamples used in the previous proofs as follows.

To impose the feasibility condition on $\mathcal{D}$, we have to compute the maximum value that $\bar{D}_i^k$

$$\bar{D}_i^k = \sum_j D_{ij}^k = \sum_k (A_{ij}^k + B_{ij}^k) \min\left\{1, \frac{\theta_0 F}{\hat{A}_j^k + \hat{B}_j^k}\right\}$$

can assume under any arrival process which is $(1, nF)$-regulated and finally impose $\bar{D}_i^k \leq SF$.

It is possible to show that the maximum value of $\bar{D}_i^k$ can be found, for the first input, considering only an arrival process in which during the $(k-l)$th observation window, only the first s_l inputs, with $1 \le s_l \le n$, receive packets destined to output $l+1$:

$$A_{i(l+1)}^{k-l} = F \quad \text{for } 1 \le i \le s_l \text{ and } 0 \le l < n.$$

According to the policy $\mathcal{D}$, we can now compute the number of packets that have to be transferred during the kth observation window, in the worst case:

$$f_l(s_l, \theta) = \max_{l \le s_1 \le k} \left\lceil \min \left\{ \frac{s_l F - \theta l F}{s_l}, \frac{\theta F}{s_l} \right\} \right\rceil^+.$$

Now the minimum speedup necessary and sufficient to transfer all the traffic is given by solving the following optimization problem:

$$\min \theta, \tag{13}$$

$$1 \le \theta \le 2, \tag{14}$$

$$\sum_{l=0}^{n-1} f_l(s_l, \theta) \le SF \tag{15}$$

for all possible $(s_0, \ldots, s_{n-1})$.

We have solved numerically the optimization problem, with an exhaustive search, and have observed that for $n < 8$, the required speedup is less than 2 and the result is useful for computing delay bounds, as shown in the next section.

5. Main results about delay performance

Under a $(1, nF)$-regulated arrival process, Theorems 3–5 evaluate the compromise between speedup and average delay penalty with respect to an OQ switch, which is $3/2 \times F$. Indeed, the average delay penalty is sum of two contributions. The first is the average delay penalty equal to F due to the F-work-conserving

Table 1
Tradeoff between speedup, the average delay penalty with respect to an OQ switch and maximum delay for a (ρ, W)-regulated traffic

Minimum speedup		Average delay penalty w.r.t. OQ	Maximum delay
Sufficient	Necessary		
$S = 1$	$S = 1$	$(3/2)\rho W$	$2\rho W$
$S = 4/3$	$S = 4/3$	$(3/4)\rho W$	$(3/2)\rho W$
$S = 3/2$	–	$(1/2)\rho W$	$(4/3)\rho W$
$S = 1.636$	–	$(3/8)\rho W$	$(5/4)\rho W$
$S = 1.739$	–	$(3/10)\rho W$	$(6/5)\rho W$
$S = 1.846$	–	$(1/4)\rho W$	$(7/6)\rho W$
$S = 1.976$	–	$(3/14)\rho W$	$(8/7)\rho W$
$S = 2$	$S = 2$	0	ρW

Table 2
Trade off between speedup, the average delay penalty with respect to an OQ switch and maximum delay for a $[\rho, \sigma]$-LBC traffic

Minimum speedup		Average delay penalty w.r.t. OQ	Maximum delay
Sufficient	Necessary		
$S = 1$	$S = 1$	$(3/2)\sigma/(1-\rho)$	$2\sigma/(1-\rho)$
$S = 4/3$	$S = 4/3$	$(3/4)\sigma/(1-\rho)$	$(3/2)\sigma/(1-\rho)$
$S = 3/2$	–	$(1/2)\sigma/(1-\rho)$	$(4/3)\sigma/(1-\rho)$
$S = 1.636$	–	$(3/8)\sigma/(1-\rho)$	$(5/4)\sigma/(1-\rho)$
$S = 1.739$	–	$(3/10)\sigma/(1-\rho)$	$(6/5)\sigma/(1-\rho)$
$S = 1.846$	–	$(1/4)\sigma/(1-\rho)$	$(7/6)\sigma/(1-\rho)$
$S = 1.976$	–	$(3/14)\sigma/(1-\rho)$	$(8/7)\sigma/(1-\rho)$
$S = 2$	$S = 2$	0	$\sigma/(1-\rho)$

property (see Theorem 1). The second is an additional average penalty equal to $F/2$ due to the switch working in a gated-fashion (see Eq. (2)). On the contrary, the absolute delay is $nF + F$, thanks to the observation at the end of Theorem 2.

Now consider an arrival process (ρ, W)-regulated and an arrival process $[\rho, \sigma]$-LBC. Tables 1 and 2 show the average delay penalty with respect to OQ and the absolute delay, for regulated and LBC traffic. Note that, for $3 < n < 8$, we computed the minimum speedup only numerically, solving the optimization problem (13)–(15). Of course, with speedup $S = 2$, a CIOQ system can emulate perfectly an OQ and the average delay penalty is null.

6. Conclusions

CIOQ switches that can control the packet delays at low speedups are very appealing. It is well known that, at speedup lower than 2, a CIOQ switch cannot emulate OQ switch even with bounded delay penalty [3]. Hence, we considered the CIOQ switch operating under a restricted, but general enough, arrival traffic class. We defined a new notion of F-work conservation for CIOQ switches, which in turn implies the property of OQ emulation with average delay penalty bounded by F. Under regulated traffic, we were able to compute an upper bound of the delay penalty for $S = 1$, $S = 4/3$ and $S = 3/2$. We computed also numerically an upper bound for some values of S, with $3/2 < S < 2$.

Thus, we showed that it is possible to emulate OQ switch under quite a general class of arrival traffic at lower speedup than 2 with bounded amount of average delay penalty.

References

[1] N. McKeown, A. Mekkittikul, V. Anantharam, J. Walrand, Achieving 100% throughput in an input-queued switch, IEEE Trans. Commun. 47 (8) (1999) 1260–1267.
[2] J. Dai, B. Prabhakar, The throughput of data switches with and without speedup, in: Proceedings of IEEE INFOCOM'00, vol. 2, Tel Aviv, Israel, March 2000, pp. 556–564.
[3] S.T. Chuang, A. Goel, N. McKeown, B. Prabhakar, Matching output queueing with a combined input/output-queued switch, IEEE J. Sel. Areas Commun. 17 (6) (1999) 1030–1039.
[4] Cisco 12000 Gigabit Switch Router, Product Overview, April 2000. http://www.cisco.com.

[5] C. Partridge, et al., A 50-Gb/s IP router, IEEE Trans. Networking 6 (3) (1998) 237–248.
[6] L. Tassiulas, Linear complexity algorithms for maximum throughput in radio networks and input queued switches, in: Proceedings of IEEE INFOCOM'98, vol. 2, New York, 1998, pp. 533–539.
[7] E. Leonardi, M. Mellia, F. Neri, M. Ajmone Marsan, Bounds on average delays and queue size averages and variances in input queued cell-based switches, in: Proceedings of IEEE INFOCOM'01, vol. 3, Anchorage, AK, April 2001, pp. 1095–1103.
[8] D. Shah, M. Kopikare, Delay bounds for the approximate maximum weight matching algorithm for input queued switches, in: Proceedings of IEEE INFOCOM'02, New York, June 2002.
[9] M. Ajmone Marsan, A. Bianco, E. Filippi, P. Giaccone, E. Leonardi, F. Neri, On the behavior of input queuing switch architectures, Eur. Trans. Telecommun. 10 (2) (1999) 111–124.
[10] P. Giaccone, B. Prabhakar, D. Shah, Towards simple, high-performance schedulers for high-aggregate bandwidth switches, in: Proceedings of IEEE INFOCOM'02, New York, June 2002.
[11] M. Ajmone Marsan, E. Leonardi, M. Mellia, F. Neri, On the stability of input-queued switches with speed-up, IEEE/ACM Trans. Networking 9 (1) (2001) 104–118.
[12] P. Krishna, N.S. Patel, A. Charny, R.J. Simcoe, On the speedup required for work-conserving crossbar switches, IEEE J. Sel. Areas Commun. 17 (6) (1999) 1057–1066.
[13] T. Weller, B. Hajek, Scheduling nonuniform traffic in a packet-switching system with small propagation delay, IEEE/ACM Trans. Networking 5 (6) (1997) 813–823.
[14] A. Borodin, J. Kleinberg, P. Raghavan, M. Sudan, D.P. Williamson, Aversarial queueing theory, J. ACM 1 (48) (2001) 13–38.
[15] J.Y. Le Boudec, P. Thiran, Network Calculus: A Theory of Deterministic Queuing Systems for the Internet, Springer, Berlin, July 2001.

Paolo Giaccone is an assistant professor at the Electronics Department of Politecnico di Torino in Italy. He obtained his Dr. Ing. and Ph.D. degrees in telecommunications engineering from Politecnico di Torino in 1998 and 2001, respectively. During the Summer 1998, he visited the High Speed Networks Research Group at Lucent Technology, Holmdel. During 2000–2001 and during the Summer 2002, he visited Prof. Balaji Prabhakar, at Electrical Engineering Department, Stanford University. Between 2001 and 2002 he held a postdoc position at Politecnico di Torino, and during the Summer 2002 at Stanford University. His main area of interest is the design of scheduling policies for high performance routers.

Emilio Leonardi is an assistant professor at the Electronics Department of Politecnico di Torino, Italy. He got a Dr. Ing. degree in electronics engineering in 1991 and a Ph.D. in telecommunications engineering in 1995 both from Politecnico di Torino. In 1995, he visited the Computer Science Department of the University of California, Los Angeles (UCLA). In the Summer 1999 he joined the High Speed Networks Research Group, at Bell Labs, while in the Summer 2001 he visited Prof. Balaji Prabhakar, at Electrical Engineering Department, Stanford University. He is co-guest editor for the special issue of IEEE JSAC on "high-performance optical/electronic switches/routers for high-speed Internet". He has co-authored over 100 papers published in international journals and presented in leading international conferences. His areas of interest are: all-optical networks, queueing theory and scheduling policies for high speed switches.

Balaji Prabhakar is an assistant professor of electrical engineering and computer science at Stanford University. He is interested in network algorithms (especially for switching, routing and bandwidth partitioning), wireless networks, web caching, network pricing, information theory and stochastic network theory. He has been a Terman Fellow at Stanford University and a Fellow of the Alfred P. Sloan Foundation. He has received the CAREER award from the National Science Foundation, the Erlang Prize from the INFORMS Applied Probability Society, and the Rollo Davidson Prize from the University of Cambridge.

Devavrat Shah is a Ph.D. candidate at the Computer Science Department, Stanford University. He received his B.Tech. from the IIT-Bombay, India in 1999 and has been at Stanford University since then. He is interested in design and analysis of network algorithms, analysis of stochastic networks and information theory. He received President of India Gold Medal from IIT-Bombay in 1999.