RZ 3425 (# 93536) 06/17/02
Computer Science 3 pages

Research Report

Work-Conservingness of CIOQ Packet Switches with Limited Output Buffers

Cyriel Minkenberg

IBM Research
Zurich Research Laboratory
8803 Rüschlikon
Switzerland

Work-Conservingness of CIOQ Packet Switches with Limited Output Buffers

Cyriel Minkenberg
IBM Research, Zurich Research Laboratory
Säumerstrasse 4, CH-8803 Rüschlikon, Switzerland
E-mail: sil@zurich.ibm.com

Abstract— **We demonstrate that no combined input- and output-queued switch with limited speed-up (i.e., smaller than the number of ports) and limited output buffering can be strictly work-conserving by constructing a counter-example traffic scenario.**

I. Introduction

Packet-switch architectures belonging to the class of *combined input- and output-queued* (CIOQ) switches (see Fig. 1) have rapidly gained popularity. The earlier approaches typically employ FIFO input buffers [1]–[6], whereas the more recent approaches have adopted *virtual output-queued* (VOQ) input buffers. In the latter category, we can distinguish between approaches with limited speed-up and centralized VOQ scheduling [7]–[13], and those with full speed-up but without centralized VOQ scheduling [14], [15]. The focus of the research in Refs. [7]–[13] has been to achieve *exact output-queuing emulation*, which means that, given identical input traffic patterns, the CIOQ switch exactly mimics the packet departures at the outputs of a reference ideal output-queued switch. The proposed algorithms such as JPM [8], LOOFA [10], and CCF [12] have been shown to be strictly work-conserving (see Def. 1) with a fabric-internal speed-up of two. However, the physically limited size of the output buffers has not yet been taken into account. Here, we will show that the property of strict work-conservingness under all traffic patterns cannot be met by a CIOQ switch with limited output buffers, regardless of speed-up.

Definition 1 (Work-Conserving) A switch that is **work-conserving** will serve any output for which at least one packet is present.

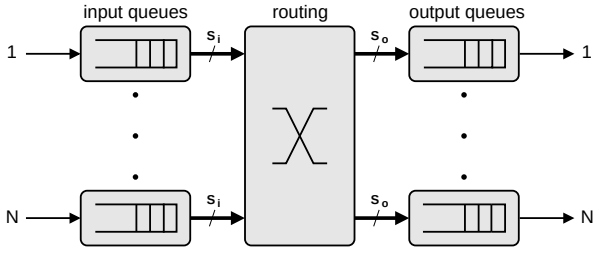

Fig. 1. Combined input and output queuing, with input speed-up S_i and output speed-up S_o. The input queues may be organized in FIFO, VOQ, or any other fashion. Arbitration is not shown here.

Figure 1 shows the logical structure of a CIOQ switch. We define the *input speed-up factor* S_i to be the number of packets that can be transmitted from a single input queue in one packet cycle (the duration of one fixed-size packet at the external line-rate), whereas the *output speed-up factor* S_o equals the number of packets that can be delivered to a single output queue in one cycle. S_i and S_o both can be between 1 and N. $S_i = 1$ and $S_o = 1$ is equivalent to classic input queuing (i.e., no output queues are needed), whereas $S_i = N$ and $S_o = N$ is equivalent to classic output queuing (no input queues are needed).[1] Given a particular speed-up S, the bandwidth of the input and/or output buffers must equal $S + 1$ times the link rate.

II. Ideal OQ emulation

A. System description

The system under study is depicted in Fig. 2, with a switch of size $N \times N$, VOQs at the input, backpressure flow control, output buffers of size Q packets each, and speed-up $S_i = S_o = S$. We assume time-slotted operation with fixed-size packets. We adopt a fabric-internal flow-control mechanism that ensures that no output queue overflows to keep the fabric lossless,[2] denoted as *backpressure mode* in [3], [4]. Backpressure is applied instantaneously if an output queue is completely full. We assume that the switch is scheduled by an ideal, centralized algorithm that can examine the status of all queues to resolve contention between inputs (at each output) and VOQs (at each input). The only restrictions we impose on this algorithm is that it must *first* guarantee work-conservingness and *second*, if possible, be fair, i.e., prevent starvation. One can easily demonstrate that fairness and strict work-conversingness are mutually exclusive [16, Sec. 3].

The question we seek to answer is whether a CIOQ packet switch as shown in Fig. 2 can emulate an ideal output-queued switch using an ideal scheduling algorithm and some limited output-queue size Q. Below, we will prove the following result: no CIOQ switch with input and output speed-up $S_i = S_o = S$, $1 \leq S < N$, and limited output buffers can be work-conserving in

[1] In case the output queues are infinite, the value of S_i is irrelevant.
[2] A lossy switch is not work-conserving by definition.

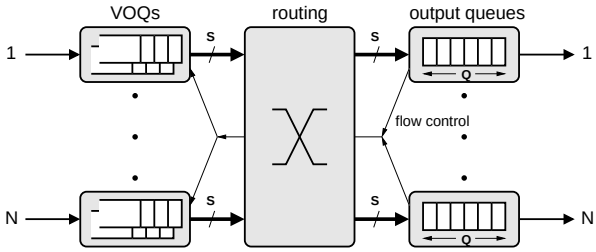

Fig. 2. CIOQ switch with VOQ, backpressure flow control, $S_i = S_o = S$, and limited output buffers.

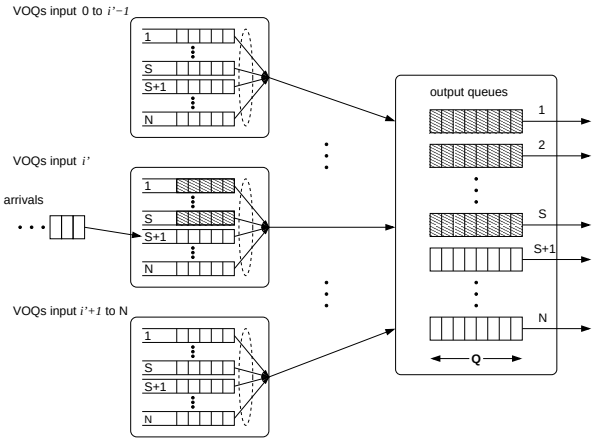

Fig. 3. Pathological traffic scenario (at t_3) to disprove ideal OQ emulation.

the sense of Definition 1 under all traffic patterns, regardless of the scheduling algorithm.[3] The proof is by counter-example. This result is important because none of the output-queued emulation schemes [7]–[13] appear to take the output buffer size into account.

B. Output contention phase

Initially, at t_0, the switch is completely empty.[4] The scenario starts with a prolonged output contention phase, during which S output queues are completely filled and substantial backlog for these outputs builds up at the inputs. We denote the arrival rate for the VOQ on input i for output j by a_i^j.

$S+1$ inputs uniformly contend for S outputs at 100% load, i.e., $a_i^j = 1/S$ for $1 \le i \le S+1, 1 \le j \le S$, and 0 otherwise. As $S+1$ packets arrive in each cycle but only S can be served from the output queues, all the outputs will be completely full by $t_1 = t_0 + QS$ at the latest. The contention continues[5] for another $m = S(S+1)\{Q(S-1)+1\}$ packet cycles. Thus, at $t_2 = t_1 + m$, there are m packets backlogged across these $S+1$ inputs. Because of the uniform traffic and the fair contention resolution, every backlogged VOQ currently has $m/(S(S+1)) = Q(S-1)+1$ packets.

From t_2 on, only input i' receives further traffic whereas all other inputs are idle. At input i' one packet arrives for each VOQ once every $S+1$ cycles, i.e., $a_{i'}^j = 1/(S+1), 1 \le j \le S$, and 0 otherwise. Note that from t_2 on all VOQs are draining at a rate of $1/(S+1)$, because $S+1$ backlogged VOQs contend for every output. As a result, the VOQs on all inputs but i' are draining at a rate of $1/(S+1)$, whereas those on input i' remain steady. Therefore, at some time $t_3 > t_2$ all inputs but i' will be completely empty.

C. Input contention phase

After the output contention phase described above, the switch system has entered a state as depicted in Fig. 3, where at t_3 exactly S output queues, say queues 1 through S, are completely full, whereas all others are completely empty. We observe the backlogged input i', which has a number of packets for each of these full output queues. From t_3 on, none of the other inputs has any traffic, so we will just look at the given input.

Assume that from t_3 on, in every cycle one packet arrives at this input destined for (empty!) output $S+1$, i.e., $a_{i'}^{S+1} = 1$. To remain work-conserving, this packet must always be served immediately, so that up to $S-1$ packets from the other VOQs can be served; we assume these are served in a round-robin (RR) order (to satisfy the fairness constraint), but note that the actual service discipline does not matter for the final result. Thus, at t_3+1, we serve one packet for output $S+1$, and one each for outputs 1 through $S-1$. At the end of the cycle, the output queues now contain Q packets at outputs 1 through $S-1$ (one arrival, one departure), and $Q-1$ packets at output S (one departure). Note that VOQs 1 through S are served in $S-1$ out of S cycles, so that their average service rate equals $(S-1)/S$. Continuing in this fashion, we find that after K cycles, with K an integer multiple of S, $K = nS, n \in \mathbb{N}$, we have served K packets for output $S+1$, and $K(S-1)/S = nS(S-1)/S = n(S-1)$ packets each for outputs 1 through S. At the same time, K packets have departed from each output queue 1 through S, so that after K cycles the occupancy of these output queues equals $Q + n(S-1) - K = Q + nS - n - nS = Q - n$ packets. Therefore, with $n = Q$, after $K = nS = QS$ cycles, all output queues will be empty. At this point, $t_4 = t_3 + QS$, there are still packets left in all of the corresponding VOQs, because at least $Q(S-1)+1$ packets were backlogged in each VOQ at t_3; now, another packet for output $S+1$ arrives. In order to remain work-conserving $S+1$ packets destined for empty output queues must be served, which is not possible because the speed-up is just S. Hence, we have demonstrated a traffic scenario under which the given switch system is not work-conserving. Figure 4 summarizes the timeline of the scenario.

As a result, no CIOQ switch system with limited output buffers of any size Q and a speed-up factor $1 \le S <$

[3]Note that work-conservingness is a necessary but not sufficient condition for exact OQ emulation.

[4]All times are expressed in packet cycles.

[5]The contention period can of course be prolonged arbitrarily to increase the backlog.

N can exactly emulate an ideal output-queued switch under all traffic patterns, regardless of the scheduling algorithm.

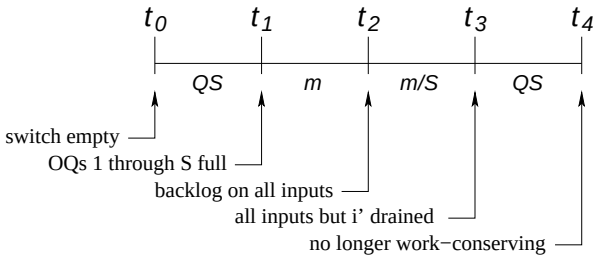

Fig. 4. Traffic scenario timeline.

The cause of this result is that, with limited buffers, the output speed-up S_o is dependent on the output-queue occupancy. The effective output speed-up S_o^j can be expressed as shown in Eq. (1):

$$S_o^j = \min(S_o, Q - q_j), \tag{1}$$

where q_j is the occupancy of output queue j.

III. Conclusion

The result derived here implies that neither the CIOQ architectures of [14]–[15], nor the OQ emulation algorithms of [7]–[13] can be practically implemented such that they are strictly work-conserving. Therefore, the output buffer size must be taken into account when studying switch performance. It also implies that more buffer space at the output is always better, so that a trade-off between the cost of additional output buffer space and the performance increase it entails must be evaluated with care. An interesting open question is whether a set of restrictions exists that, when imposed on the offered traffic, guarantees strict work-conservingness for a given output buffer size.

Acknowledgment

The author thanks Mitchell Gusat for helpful discussions on the traffic scenario.

References

[1] Y. Oie, M.Murata, K. Kubota and H. Miyahara, "Effect of Speedup in Nonblocking Packet Switch," in *Proc. ICC '89*, Jun. 1989, pp. 410-414.

[2] A.K. Gupta and N.D. Georganas, "Analysis of a Packet Switch with Input and Output Buffers and Speed Constraints," in *Proc. IEEE INFOCOM '91*, Bal Harbour, FL, Apr. 1991, pp. 694-700.

[3] I. Iliadis and W.E. Denzel, "Analysis of Packet Switches with Input and Output Queuing," *IEEE Trans. Commun.*, vol. 41, no. 5, May 1993, pp. 731-740.

[4] A. Pattavina and G. Bruzzi, "Analysis of Input and Output Queueing for Nonblocking ATM Switches," *IEEE/ACM Trans. Netw.*, vo. 1, no. 3, Jun. 1993, pp. 314-328.

[5] C.-Y. Chang, A.J. Paulraj and T. Kailath, "A Broadband Packet Switch Architecture with Input and Output Queueing," in *Proc. IEEE GLOBECOM '94*, pp. 448-452.

[6] M.J. Lee and D.S. Ahn, "Cell Loss Analysis and Design Trade-Offs of Nonblocking ATM Switches with Nonuniform Traffic," *IEEE/ACM Trans. Netw.*, vol. 3, no. 2, Apr. 1995, pp. 199-210.

[7] N. McKeown, B. Prabhakar and M. Zhu, "Matching Output Queueing with Combined Input and Output Queueing," in *Proc. 35th Annual Allerton Conf. Communication, Control and Computing*, Monticello, IL, Oct. 1997.

[8] I. Stoica and H. Zhang, "Exact Emulation of an Output Queueing Switch by a Combined Input Output Queueing Switch," in *Proc. 6th IEEE/IFIP IWQoS '98*, Napa Valley, CA, May 1998, pp. 218-224.

[9] A. Charny, P. Krishna, N. Patel and R.J. Simcoe, "Algorithms for Providing Bandwidth and Delay Guarantees in Input-Buffered Crossbar Switches with Speedup," in *Proc. IWQoS '98*, Napa Valley CA, May 1998, pp. 225-234.

[10] P. Krishna, N.S. Patel, A. Charny and R.J. Simcoe, "On the Speedup Required for Work-Conserving Crossbar Switches," *IEEE J. Sel. Areas Commun.*, vol. 17, no. 6, 1999, pp. 1057-1066.

[11] B. Prabhakar and N. McKeown, "On the Speedup Required for Combined Input and Output Queued Switching," *Automatica*, vol. 35, 1999.

[12] S.-T. Chuang, A. Goel, N. McKeown and B. Prabhakar, "Matching Output Queueing with a Combined Input Output Queued Switch," *IEEE J. Sel. Areas Commun.*, vol. 17, no. 6, Jun. 1999, pp. 1030-1039.

[13] J.G. Dai and B. Prabhakar, "The Throughput of Data Switches with and without Speedup," in *Proc. INFOCOM 2000*, Tel Aviv, Israel, Mar. 2000, vol. 2, pp. 556-564.

[14] R. Fan, H. Suzuki, K. Yamada, and N. Matsuura, "Expandable ATOM Switch Architecture (XATOM) for ATM LANs," in *Proc. ICC '94*, vol. 1, New Orleans, LA, May 1994, pp. 402-409.

[15] F.M. Chiussi and A. Francini, "A Distributed Scheduling Architceture for Scalable Packet Switches," *IEEE J. Sel. Areas Commun.*, vol. 18, no. 12, Dec. 2000, pp. 2665-2683.

[16] N. McKeown, V. Anantharam and J. Walrand, "Achieving 100% Throughput in an Input-Queued Switch," in *Proc. INFOCOM '96*, San Francisco, CA, Mar. 1996, vol. 1, pp. 296-302.

[17] M. Katevenis, D. Serpanos and E. Spyridakis, "Switching Fabrics with Internal Backpressure using the ATLAS I Single-Chip ATM Switch," in *Proc. GLOBECOM '97*, Phoenix, AZ, Nov. 1997.

[18] C. Minkenberg and T. Engbersen, "A Combined Input- and Output-Queued Packet-Switch System Based on PRIZMA Switch-on-a-Chip Technology," *IEEE Commun. Mag.*, vol. 38, no. 12, Dec. 2000, pp. 70-77.