Load-Balanced Combined Input-Crosspoint Buffered Packet Switch and Long Round-Trip Times

Roberto Rojas-Cessa, Ziqian Dong, and Zhen Guo

***Abstract*— The amount of memory in buffered crossbars is proportional to the number of crosspoints, or $O(N^2)$, where N is the number of ports, and to the crosspoint buffer size, which is defined by the distance between the line cards and the buffered crossbar, to achieve 100% throughput under high-speed data flows. A long distance between these two components can make a buffered crossbar costly to implement. In this letter, we propose a load-balanced combined input-crosspoint buffered packet switch that uses small crosspoint buffers and no speedup. The proposed switch reduces the required size of the crosspoint buffers by a factor of N and keeps the cells in sequence.**

***Index Terms*— Buffered crossbar, round-trip time, crosspoint buffer, Birkhoff-Von Neumann, load balancing**

I. Introduction

As optical technologies spread quickly and ubiquitously, it is becoming feasible to transmit single flows with increasingly high data rates. High-performance switches and routers are required to handle such flows and, therefore, to provide high-speed ports.

Combined input-crosspoint buffered (CICB) switches are an alternative to input-buffered switches to relax arbitration timing and to provide high-performance switching for packet switches with high-speed ports [1].[1] Incoming variable-size packets are segmented into fixed-length packets, called cells, at the ingress side of a switch and re-assembled at the egress side, before the packets depart from the switch. This letter considers the use of cells.

The amount of memory in a buffered crossbar is $N^2 \times k \times L$, where N is the number of input/output ports, k is the crosspoint buffer size in number of cells, and L is the cell size in bytes. The value of k is defined by the length of the round-trip time (RTT), defined in [2] as the sum of the delays of 1) the input arbitration IA, 2) the transmission of a cell from an input to the crossbar $d1$, 3) the output arbitration OA, and 4) the transmission of the flow-control information back from the crossbar to the input, $d2$. Cell and bit alignments are included in the transmission times. For example, the switch proposed in [2] requires the size of k be equal to or larger than the round-trip time to avoid throughput degradation or crosspoint-buffer underflow for flows (here defined as the data arriving at input i and destined to output j, where $0 \leq i, j \leq N-1$) with high data rates.

In a CICB switch, the crosspoint-buffer size to avoid underflow by flows of data rate C b/s, where C is the port speed, is $RTT = d1+OA+d2+IA \leq k$, such that cells are transmitted continuously every time slot [2].

Furthermore, as the buffered crossbar can be physically located far from the input ports, actual RTTs can be long. To support long RTTs in a buffered-crossbar switch, the crosspoint-buffer size needs to be increased [3], such that up to RTT cells can be buffered. However, the memory amount that can be allocated in a chip is limited, and therefore, it can make the implementation costly or infeasible when the distance between line cards and the buffered crossbar is long. An interesting scheme using limited memory is presented in [4] for a switch with p traffic classes, where the crosspoint buffer size is larger than RTT for a single class, and smaller than $p \times RTT$.

A solution to keep the crosspoint buffer small while supporting long RTTs and high data rates is needed. In this paper, we study a CICB switch that uses round-robin arbitration and credit-based flow control under long round-trip times and high data-rate flows. We show the throughput degradation as a function of the round-trip time and the crosspoint buffer size in buffered crossbar switches with dedicated connections and crosspoints. By considering the Birkhoff-Von Neumann switch [5], we propose using an extra switching stage, as a load balancer, with a buffered crossbar to relax the crosspoint buffer size such that flows with high data rates can be handled when $k < RTT$. We show that this architecture supports high data-rate flows and $RTT = kN$, which results in a crosspoint-buffer of size $\frac{1}{N}$ of that in a switch without load-balancing stage.

This letter is organized as follows. Section II discusses the effect of long round-trip times in a CICB switch. Section III introduces the proposed load-balanced CICB switch. Section IV presents the throughput performance of the load-balanced CICB switch. Section V presents the conclusions.

II. Effects of Long Round-Trip Time and Limited k

To keep up with high data rates, switch ports must be able to handle flows of up to C b/s,[2] where C is the data-rate capacity of a port in a switch or router. In a CICB switch (e.g., the

This work is supported in part by National Science Foundation under Grants 0435250 and 0423305.

The authors are with the Department of Computer and Electrical Engineering, New Jersey Institute of Technology, University Heights, Newark NJ 07102. Roberto Rojas-Cessa is the correspondence author. Email: rrojas@njit.edu.

[1]There is a large number of works in CICB switches that cannot be cited here for lack of space.

[2]In contrast, switches unable to support such flows can only handle aggregated data rates of C b/s, where each flow might have a data rate r_{single}, such that $r_{single} < C$.

CIXB switch presented in [2]), the maximum flow rate that can be supported is $C\frac{k}{RTT}$. Note that when $r_{f(i,j)} = C$, where $r_{f(i,j)}$ is the rate of $f(i,j)$, the maximum flow rate that the CIXB switch can transfer from inputs to outputs is equivalent to its achievable throughput.

We simulated the CIXB switch to observe the throughput obtained under different k and RTT values in a 32×32 switch, and to validate the traffic model to test the proposed architecture. Different from [2], we consider $RTT > 0$ in this letter. Here, we assume that the distances between input ports and the buffered crossbar are identical.[3] To model flows with different rates, we consider the unbalanced traffic model [2]. This model uses $w + \frac{1-w}{N}$ as the fraction of the input load directed from input i to output j, for $j = i$, where w is the unbalanced probability, and the fraction $\frac{1-w}{N}$ of the input load directed from input i to output j for $j \neq i$ (i.e., uniform distribution). Therefore, the fraction of C that $f(i,j)$ uses is $r_{f(i,j)} = w + \frac{1-w}{N}$. The maximum data rate of $f(i,j)$ is represented by setting $w = 1$ or $r^{max}_{f(i,j)} = C$, and the minimum data rate is represented when $w = 0$ or $r^{min}_{f(i,j)} = \frac{1}{N}$. We emphasize our observations in these two w values in the unbalanced traffic model.

Figure 1 shows that flows with a rate $r_{f(i,j)} = r^{min}_{f(i,j)}$ (i.e., w=0), the throughput is 100% for $k = RTT$, as shown by curves 1) and 5), and for longer RTTs or $k < RTT$, as shown by curves 3)-6), where $k \geq 2$. However, the throughput is less than 100%, as shown by curve 2), where $RTT = 31$ and $k = 1$. The low data rate and uniform distribution of traffic relax the demand for crosspoint buffer space, resulting in high throughput.

However, as the data rate of the flow increases (i.e., w), throughput degradation increases as RTT becomes longer. The worst-case scenario is observed when $r_{f(i,j)} = C$ b/s (i.e., w=1) where the achieved throughput is $\frac{k}{RTT}$ for $RTT > k$, as curves 2), 3), 4), and 6) show.

III. Load-Balanced Combined Input-Crosspoint Buffered Switch

The switch has virtual output queues (VOQs) in the input ports, a load-balancing stage (e.g., bufferless crossbar), and a buffered crossbar. In this switch, input ports are also called external inputs, each of which is denoted as EI_i, at the load-balancer side. The outputs of the load-balancing stage are called internal outputs, each of which is denoted as IO_h, where $0 \leq h \leq N-1$. IOs are physically equivalent to the inputs of the buffered crossbar, also called internal inputs, each of which is denoted as II_h. The outputs of the buffered crossbar, or output ports, are also called external outputs, each of which is denoted as EO_j.

The load-balancing stage uses pre-determined and cyclic configurations connecting its inputs and outputs at predetermined time slots (e.g., EI_i is connected to $IO_{(h+t) \bmod N}$, where t is any given time slot), and a buffered crossbar. A crosspoint in the buffered crossbar that connects II_h to EO_j, is denoted as $XP(h,j)$. The buffer at $XP(h,j)$ is denoted as $XPB(h,j)$. Figure 2 shows this architecture, where the transmission delays between ports and the crosspoint are denoted by $d1$ and $d2$. We consider crosspoint buffers where $k \geq 1$. There are N VOQs at each input. A VOQ at input i that stores cells for output j is denoted as $VOQ(i,j)$. Each EI_i has N VOQ counters $VC_{i,j}$, that counts the number of cells in $VOQ_{i,j}$ that have not been selected for dispatching.

At EI_i, cells destined to output j arrive at $VOQ(i,j)$ and wait for dispatching until they are selected by the input arbiter. The input arbiter, placed in the buffered crossbar, selects the next cell to be forwarded to the crossbar. The input arbiter uses a round-robin schedule to select a non-empty VOQ by considering those $VC_{h,j} > 0$. When a cell is dispatched by the input, the packet traverses the load-balancing stage, which follows a pre-determined and fixed order to connect EIs to IOs. A cell going from EI_i to EO_j may enter the buffered crossbar through II_h and be stored in $XPB(h,j)$. Cells leave EO_j after being selected by the output arbiter. The output arbiter uses first-come first-serve (FCFS) selection to keep cells of $f(i,j)$ in sequence. The output arbiter considers the time when a cell arrives at the crosspoint buffer to perform FCFS among dedicated crosspoint buffers. We use the following theorem to explain why FCFS, used as a selection policy by output arbiters, is sufficient to keep cells in sequence.

Theorem 1: Cells are served in-sequence when First-Come First-Serve (FCFS) is used as selection policy by an output arbiter in a load-balanced CICB switch, for any XPB size.

For the sake of brevity, we summarize the proof of the theorem as follows.

Proof: Since one cell is transferred from $VOQ_{i,j}$ to $XPB_{i,j}$ per time slot, and there are no buffers in between, then cells from that VOQ (or flow) arrive in the departure order. As the crossbar assigns a sequence service number (SSN), which is equivalent to the arrival time, to each cell at arrival then cells in a crosspoint buffer are placed in the XPB in the order they arrived the buffered crossbar as the management of cells in an XPB follows a First-In First-Out policy.

Because an output arbiter uses FCFS selection, the SSN of a cell determines the service order, independently of the contents in other XPBs. Therefore, it is clear that, $\forall\, c1_{i,j}(t_a)$, or a cell from input i destined to output j that arrived at t_a to XPB of output j, there is no $c2_{i,j}(t_b)$, where $t_b > t_a$, such that $c2$ departs from output j before cell $c1$. This is feasible because all crosspoint buffers can be located within the same chip and the assigment of SSN is easy to implement. ■

The input arbitration is performed by considering $VC_{h,j}$ and crosspoint occupancy. The selection information is sent from the buffered crossbar to the corresponding VOQ (i.e., $d2$). Cells and VOQ selection information experience transmission delay between input ports and the buffered crossbar.

[3]The results in this letter also apply for non-identical distances.

IV. Throughput of the Load-Balanced CICB Switch

We observe the effect of long RTTs in the proposed switch model by measuring the switch throughput under the unbalanced traffic model, as discussed in Section II.

Figure 3 shows the throughput performance of the load-balanced CICB switch, with $k \geq 1$. The switch achieves 100% throughput when $kN - RTT \geq 0$ and $r_{f(i,j)} = r^{max}_{f(i,j)} = r^{min}_{f(i,j)}$ (or $w = 0$ and $w = 1$) as the figure shows. These results show that long RTTs and flows with high data rates can be supported by this switch.

The decreased throughput around w=0.7 in the curves where $kN - RTT$ is small or close to 0, is the result of having a limited and small k, mixed traffic (the high data-rate flows are mixed with a large number of low data-rate flows) as described in Section II, and round-robin arbitration at the inputs. Nevertheless, this throughput is higher than that of the switch in Section II. In these cases, a more effective arbitration scheme [6] can be used to improve the throughput for small $kN - RTT$ values. Note that the throughput is close to 100% for all w and for large $kN - RTT$ values. In general, the load-balancing stage improves the switching performance.

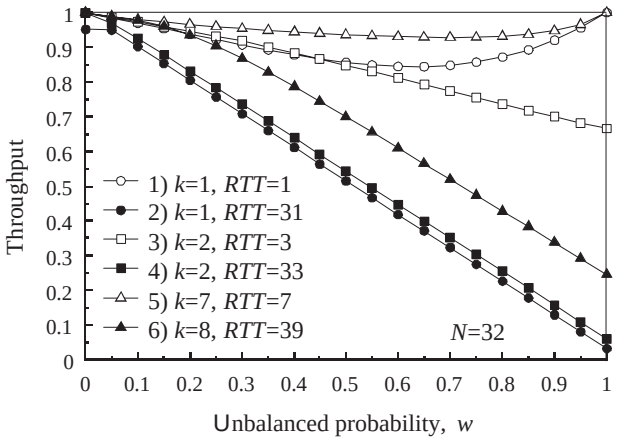

Fig. 1. Throughput performance of a CICB switch [2] with $RTT > 0$.

V. Conclusions

We presented the effect of long round-trip times RTTs, where the crosspoint buffer size k is such that $k < RTT$. We observed that switches based on buffered crossbars with dedicated crosspoint-buffer access have their maximum throughput as the ratio of $\frac{k}{RTT}$, when input ports handle a single flow with a data rate equal to the port capacity. To minimize the crosspoint-buffer size, we proposed a switch model that uses a load-balancing stage in front of the buffered crossbar, such inputs can flexibly access different crosspoint buffers. The proposed switch supports RTTs that can be kN-time-slot long, while providing 100% throughput for such high data-rate flows. As a comparison, for a given RTT size, the load-balanced CICB switch requires a minimum $k = \frac{RTT}{N}$ cells while a simple CICB switch requires a minimum $k = RTT$ cells. Therefore, the proposed switch relaxes the amount of memory to $\frac{1}{N}$ of the amount required by a CICB switch with dedicated access to crosspoint buffers.

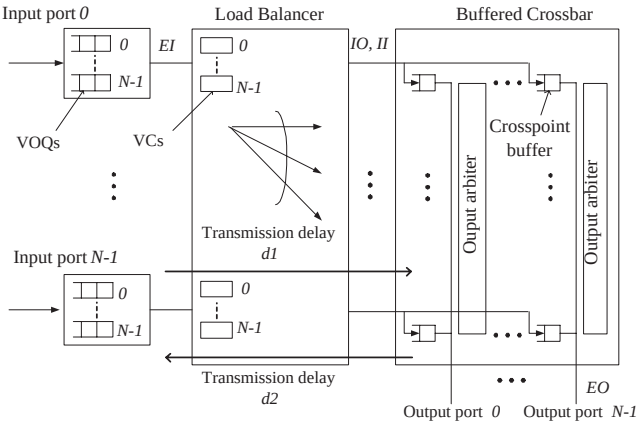

Fig. 2. $N \times N$ buffered crossbar with a load-balancing stage.

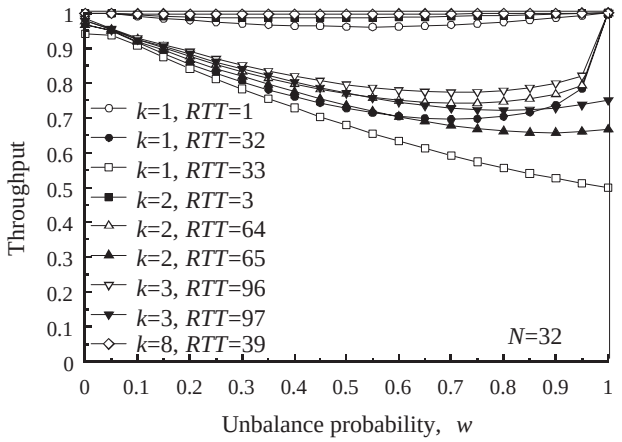

Fig. 3. Throughput of the 32×32 load-balanced CICB switch.

References

[1] Y. Doi and N. Yamanaka, "A High-Speed ATM Switch with Input and Cross-Point Buffers," *IEICE Trans. Commun.*, vol. E76, no.3, pp. 310-314, March 1993.

[2] R. Rojas-Cessa, E. Oki, Z. Jing, and H. J. Chao, "CIXB-1: Combined Input-One-Cell-Crosspoint Buffered Switch," *Proceedings of IEEE HPSR 2001*, pp. 324-329, May 2001.

[3] F. Abel, C. Minkenberg, R. P. Luijten, M. Gusat, and I. Iliadis, "A Four-Terabit Single-Stage Packet Switch with Large Round-Trip Time Support," *Proceedings of Hot Interconnects, 2002. 10th Symposium on*, pp. 5-14, Aug. 2002.

[4] R. Luijten, C. Minkenberg, and M. Gusat, "Reducing Memory Size in Buffered Crossbars with Large Internal Flow Control Latency," *Proceedings of IEEE Global Telecommunications Conference 2003*, Vol. 7, pp. 3683-3687, Dec. 2003

[5] C-S. Chang, D-S. Lee, and Y-S. Jou, "Load Balanced Birkhoff-Von Neumann Switches," *Proceedings of IEEE HPSR 2001*, pp. 276-280, May 2001.

[6] R. Rojas-Cessa and E. Oki, "Round-Robin Selection with Adaptable-Size Frame in a Combined Input-Crosspoint Buffered Switch," *IEEE Commun. Letters*, vol. 7, issue 11, pp. 555-557, November 2003.