

# Buffered Crossbar (CICQ) Switch Architecture

[Manolis Katevenis](#), [Nikolaos Chrysos](#), [Georgios Passas](#), and [Dimitrios Simos](#);

with the cooperation of

[Dionysios Pnevmatikatos](#), Ioannis Papaefstathiou, and Georgios Kalokerinos.

[Computer Architecture and VLSI Systems \(CARV\) Lab](#),

[Institute of Computer Science \(ICS\), FORTH](#)

Science and Technology Park of Crete, P.O.Box 1385, Heraklion, Crete, GR-711-10 Greece

© copyright 2003-2004 by FORTH and IEEE

## OUTLINE:

The crossbar is the most frequently used switching element topology. It offers simplicity and non-blocking operation. However, *when bufferless*, it also requires a centralized scheduler, which must simultaneously satisfy --in each cell time-- all input and all output link constraints. The cost and complexity of this scheduler increases considerably for short cell times and for large switch sizes; additionally, these schedulers cannot practically offer WFQ-type QoS. Furthermore, bufferless crossbars can only be efficiently used with fixed-size cells arriving from mutually-synchronized line cards; when we need to switch variable-size packets, we must first segment them into fixed-size cells. To compensate for the inefficiencies of scheduling and of packet segmentation, internal (crossbar) speedup is used; commercial crossbars often use a speedup factor of 2 to 3. The net effect is to limit the maximum external line rate to roughly one half to one third the peak achievable crossbar line rate.

The operation of the crossbar can be dramatically improved by including *small buffers at each crosspoint*; CMOS technology has recently reached the point where this is feasible for the buffer sizes that are needed in order for backpressure flow control to operate efficiently between the crossbar and the VOQ's in the ingress line cards. This "*buffered crossbar*" or "*combined input-crosspoint queueing (CICQ)*" architecture has significant advantages over the previous, traditional bufferless configuration:

- i. The scheduling task is dramatically simplified; WFQ-type QoS is easily implementable; there are no scheduler inefficiencies to be compensated by speedup.
- ii. The crossbar can operate directly on variable-size packets, hence there is no need for segmentation and reassembly circuits; the need for mutually synchronized line cards (at the cell-time level) is also eliminated.
- iii. Internal speedup is not needed, because there is no packet segmentation and no scheduler inefficiencies; hence, the external line rate can be as high as the crossbar line rate.
- iv. The egress path of the switch needs no buffer memory --at least no large, off-chip memory--

because packet reassembly is not needed, and because, in the lack of internal speedup, there is no output queue build up; this eliminates a major cost component.



*In a bufferless crossbar, the scheduling decisions at the input and output ports all depend on each other: each output can only be paired to a single input and conversely for the inputs*



*Small buffer memories at the crosspoints allow distributed scheduling decisions; operation with variable-size packets now becomes feasible*

For an introductory explanation page, for the non-specialist, [click here](#).

We have studied WFQ-type scheduling in cell-based CICQ switches, and we are currently studying the design and detailed operation of buffered crossbars operating directly on variable-size packets:

## 1. Distributed (WFQ) Scheduling in Buffered Crossbars

The scheduling task is dramatically simplified in buffered crossbars: distinct servers at each input and each output collectively but still independently schedule the set of flows through the interconnect; they are loosely coordinated through backpressure signals from the crosspoint buffers. We have analyzed such *distributed scheduling policies* in buffered crossbars operating on fixed-size cells and using *weighted fair queueing* (WFQ) schedulers at each input and output.

- Our results are presented in various **papers (2002-2003)**, available through another [page](#).

## 2. Variable-Packet-Size Buffered Crossbars:

- M. Katevenis, G. Passas, D. Simos, I. Papaefstathiou, N. Chrysos: "Variable Packet Size Buffered Crossbar (CICQ) Switches", *Proc. IEEE International Conference on Communications (ICC 2004)*, Paris, France, 20-24 June 2004, vol. 2, pp.1090-1096.
  - Preprint in [PDF](#) (250 KBytes) or [Postscript](#) (540 KBytes); © Copyright 2004 by IEEE.
  - Talk Transparencies in [PPT](#) (290 KBytes) or [PDF](#) (205 KBytes); © Copyright 2004 by FORTH.

**ABSTRACT:**One of the most widely used architectures for packet switches is the crossbar. A special version of it is the buffered crossbar, where small buffers are associated with the crosspoints; this simplifies scheduling and improves its efficiency and QoS capabilities to the point where the switch needs no internal speedup. Furthermore, by supporting variable length packets throughout a buffered crossbar: (a) there is no need for segmentation and reassembly (SAR) circuits; (b) no speedup is necessary to support SAR; and (c) synchronization between the input and output clock domains is simplified. In turn, the lack of SAR and speedup mean that no output queues are needed, either. In this paper we present an architecture, a chip layout and cost analysis, and a performance evaluation of such a 300 Gbps buffered crossbar operating on

variable-size packets. The proposed organization is simple yet powerful, can be implemented using modern technology, and, as the performance results demonstrate, it clearly outperforms unbuffered crossbars.

[*Previous, outdated* version: Sep. 2003, 8 pages, in [pdf](#) (140 KB) or [ps](#) (310 KB)].

- D. Simos: "Design of a 32x32 Variable-Packet-Size Buffered Crossbar Switch Chip", *Technical Report FORTH-ICS/TR-339*, Inst. of Computer Science, FORTH, Heraklion, Crete, Greece; M.Sc. Thesis, Univ. of Crete; July 2004, 102 pages.  
- Available in [PDF](#) (1.15 MBytes) format; © Copyright 2004 FORTH.

This technical report gives the details of the design of the chip described in the above paper "Variable Packet Size Buffered Crossbar (CICQ) Switches". In particular, we present the design, using a hierarchical ASIC flow, of a 32x32 buffered crossbar chip core, operating directly on variable-size packets using a 2 KByte 2-port SRAM buffer in each of the 1024 crosspoints, and providing 300 Gb/s of aggregate bandwidth in 0.18-micron CMOS technology. In this technology, core area is 420 square mm, and core power is 6 W; extrapolations for 0.13-micron CMOS indicate an estimated core area of 200 square mm, and core power of 3.2 W. The majority of core power is consumed in driving cross-chip wires, while memories and logic are minority consumers. Hierarchical ASIC flows are difficult to use, but became necessary due to the large size of the design. We present the detailed system design (block diagrams as well as critical circuit details), followed by a description of the design flow, including its numerous intricacies and the lessons that we learnt. In particular, we describe the choice of a hierarchy that is appropriate for effective placement, routing, and timing behavior. The final placement and routing showed that the synthesis tool had underestimated the design area by 30%, due to the dominance of long (end-to-end) wires in this design.

- G. Passas: "Performance Evaluation of Variable Packet Size Buffered Crossbar Switches", *Technical Report FORTH-ICS/TR-328*, Inst. of Computer Science, FORTH, Heraklion, Crete, Greece; B.Sc. Thesis, Univ. of Crete; November 2003, 46 pages.  
- Available in [PDF](#) (350 KBytes) or [Postscript](#) (1.2 MBytes) format; © Copyright 2003 FORTH.

This technical report describes in more detail the simulator used for the performance evaluation in the above paper "Variable Packet Size Buffered Crossbar (CICQ) Switches". It also contains additional simulation results.

### 3. Multiple Priority Levels in Buffered Crossbars:

- N. Chrysos, M. Katevenis: "Multiple Priorities in a Two-Lane Buffered Crossbar", *Inst. of Computer Science, FORTH, Heraklion, Crete, Greece*, March 2004, 7 pages.  
- Available in [PDF](#) (285 KBytes) or [Postscript](#) (330 KBytes) format; © Copyright 2004 FORTH.

**ABSTRACT:** A significant advantage of buffered crossbar (combined input-crosspoint queueing - CICQ) switches is that they can directly operate on variable-size packets, thus saving the costs and inefficiencies of packet segmentation and reassembly (SAR). However, in order to support multiple priority levels, separate queues per priority are needed at each crosspoint, in order to prevent HOL blocking and buffer hogging; these queues are expensive because they each need a size of at least one maximum-size packet. In this paper we propose a scheme that uses only two queues per crosspoint to effectively support multiple priorities. We adaptively adjust the priority levels of the two queues so that most traffic goes through the "lower" queue, while the "upper" queue remains usually available for higher priority packets to overtake the former. Through simulation, and assuming 8 priority levels, we compare our scheme to an ideal system that uses 8 queues per crosspoint. For realistic traffic, the two systems perform almost identically, although ours uses 4 times less memory in the crossbar. Even under a highly irregular traffic pattern Bursts60, our system will not increase the average delay of any priority level by more than 75 percent compared to the ideal system.

[*Previous, outdated* version: Sep. 2003, 8 pages, in [pdf](#) (300 KB) or [ps](#) (400 KB)].

- N. Chrysos: "Design Issues of Variable-Packet-Size, Multiple-Priority Buffered Crossbars", Technical Report FORTH-ICS/TR-325, *Inst. of Computer Science, FORTH, Heraklion, Crete, Greece*, October 2003, 32 pages.  
- Available in [PDF](#) (850 KBytes) or [Postscript](#) (1.4 MBytes) format; © Copyright 2003 FORTH.

This technical report describes in more detail the methods proposed in the above paper "Multiple Priorities in a Two Lane Buffered Crossbar". It also contains additional simulation results. Further on, it presents the RS method, which reduces the complexity in the ingress line-cards and also reduces the HOL blocking and buffer hogging behavior. Finally, it considers several design issues related to variable-packet-size buffered crossbars: alternative positions for the input schedulers, storage of credits within the credit chip, scheduling of operations at the contention points, cut-through, store-and-forwarding, and crosspoint buffers dimensioning.

### **Acknowledgements:**

Financial support was provided partly through project 002075 "SIVSS" of the European Union FP6 IST Programme. The CAD tools for chip design were provided by the University of Crete, through Europractice. Georgios Sapountzis helped us shape our ideas; we deeply thank him. We also acknowledge the assistance of C. Georgis.

---

### **© Copyright 2003-2004 by IEEE or FORTH:**

These papers are protected by copyright. Permission to make digital/hard copies of all or part of this material without fee is granted provided that the copies are made for personal use, they are not made or distributed for profit or commercial advantage, the IEEE or FORTH copyright notice, the title of the publication and its date appear, and notice is given that copying is by permission of the IEEE or of the

Foundation for Research & Technology -- Hellas (FORTH), as appropriate. To copy otherwise, in whole or in part, to republish, to post on servers, or to redistribute to lists, requires prior specific written permission and/or a fee.

---

[Up to Packet Switch Architecture R&D at CARV-ICS-FORTH](#)

Last updated: July 2004, by [M. Katevenis](#).