

Distributed Scheduling in Buffered Crossbars (CICQ)

Nikolaos Chrysos, George F. Georgakopoulos, [Manolis Katevenis](#)

[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
and

[Dept. of Computer Science, University of Crete](#), Heraklion, Crete, Greece

© copyright 2002-2003 by FORTH and the Univ. of Crete, Greece

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 with switch size. Moreover, it is an open problem whether and how such schedulers can efficiently implement sophisticated policies like weighted-fair-queueing (WFQ) / weighted-round-robin (WRR). Also, bufferless crossbars can only be efficiently used with fixed-size cells arriving from mutually-synchronized line cards.

Modern technology allows small buffers (on the order of a few cells) to be included at each crosspoint of a crossbar. In such a *buffered crossbar*, the scheduling task is dramatically simplified: 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. Also, buffered crossbars do not require synchronized line cards, and can operate efficiently even with variable-size packets (our next research topic):

[[Up to Buffered Crossbar \(CICQ\) Switch Architecture at FORTH](#)]

We have analyzed such *distributed scheduling policies* in buffered crossbars using *weighted fair queueing* (WFQ) schedulers at each input and output. Initially, we assumed fixed-size cell traffic. We found, using extensive simulations and analysis, that this system approximates very well the ideal *weighted max-min fair* allocation of throughput. We also studied the convergence time, and we quantified the unfairness during the convergence process. Further, we studied saturation throughput under various assumptions on the form of traffic. Crosspoint buffers of size just 1 cell each suffice for the scheduling operation; crosspoint buffers of size 4 to 5 cells each yield excellent performance, at least for switches up to 32x32.

1. Introductory Paper: simulation results on WMMF allocation

N. Chrysos, M. Katevenis: "Weighted Fairness in Buffered Crossbar Scheduling", Proceedings of the IEEE Workshop on *High Performance Switching and Routing (HPSR 2003)*, Torino, Italy, June 2003, pp. 17-22.

Available in [PDF](#) (120 KBytes) or [Postscript](#) (180 KBytes) or [gzip'ed Postscript](#) (50 KBytes) format; © Copyright 2003 IEEE. The slides of the presentation are also available, in [PDF](#) format (220 KBytes).

ABSTRACT:The crossbar is the most popular packet switch architecture. By adding small buffers at the crosspoints, important advantages can be obtained: (1) Crossbar scheduling is simplified. (2) High throughput is achievable. (3) Weighted scheduling becomes feasible. In this paper we study the fairness properties of a buffered crossbar with weighted fair schedulers. We show by means of simulation that, under heavy demand, the system will allocate throughput in a weighted max-min fair manner. We study the impact of the size of the crosspoint buffers in approximating the weighted max-min fair rates and we find that a small amount of buffering per crosspoint (3-8 cells) suffices for the maximum percentage discrepancy to fall below 5% for 32x32 switches.

2. Analysis Papers: WMMF allocation

G. Georgakopoulos: "Nash Equilibria as a Fundamental Issue Concerning Network-Switches Design", *Proc. IEEE International Conference on Communications (ICC 2004)*, Paris, France, 20-24 June 2004, vol. 2, pp. 1080-1084.

- Preprint in [PDF](#) (200 KBytes); © Copyright 2004 by IEEE.

ABSTRACT:We view the 'packet-switching problem' (from N inputs towards N outputs) from the perspective of game theory and we prove that, if the rates of flows are weighed, then 'weighed max-min fair service rates' are the unique Nash equilibrium point of a natural strategic game in which throughput is granted on a 'least-demanding first-served' principle. We prove that a crossbar switching device with suitably randomized schedulers converges to this equilibrium point without pre-computing it.

[*Previous, outdated* version: June 2003, 12 pages, in [pdf](#) (220 KBytes)].

G. Georgakopoulos: "Few buffers suffice: Explaining why and how crossbars with weighted fair queuing converge to weighted max-min fairness", *Dept. of Computer Science, University of Crete, Heraklion, Crete, Greece*, July 2003, 6 pages.

Available in [PDF](#) format (240 KBytes); © Copyright 2003 University of Crete.

ABSTRACT:We explain why and how a crossbar switch with weighted fair queues (a

deterministic device) converges to weighted max-min fair rates of service. Based on a plausible technical hypothesis we prove that a small mean number of buffers suffices to this end. WMMF-rates comprise the unique Nash equilibrium of a natural switching game --a fact that supports further their use.

3. Transient Behavior Papers: simulation results

N. Chrysos, M. Katevenis: "Transient Behavior of a Buffered Crossbar Converging to Weighted Max-Min Fairness", *Inst. of Computer Science, FORTH, Heraklion, Crete, Greece*, August 2002, 13 pages. Available in [PDF](#) (540 KBytes) or [Postscript](#) (585 KBytes) or [gzip'ed Postscript](#) (145 KBytes) format; © Copyright 2002 FORTH.

In addition to the HPSR-2003 paper (section 1 above), this paper provides arguments for why the system converges to weighted max-min (WMM) fairness, and studies the factors that affect stabilization delay after changes in offered load or weight factors. Transient behavior simulations verified that stabilization delay is proportional to buffer size, and inversely proportional to the magnitude of the change in bandwidth allocation.

N. Chrysos: "Weighted Max-Min Fairness in a Buffered Crossbar Switch with Distributed WFQ Schedulers: a First Report", *Technical Report FORTH-ICS/TR-309*, Inst. of Computer Science, FORTH, Heraklio, Crete, Greece; M.Sc. Thesis, Univ. of Crete (advisor: M. Katevenis); April 2002, 150 pages. Available in [Postscript](#) (5.4 MBytes) or [gzip'ed Postscript](#) (1.4 MBytes) format; © Copyright 2002 FORTH.

The M.Sc. Thesis **in Greek** (81 pages) is also available, in [Postscript](#) (1.4 MBytes) format; © Copyright 2002 Univ. of Crete.

This 150-page thesis contained lots of intermediate results of our research; see in particular section 5.3 (pp. 77-93) on the *transient behavior* when flow eligibility (or weights) change. The Greek version is a compact collection of the most important results from the English TR.

Acknowledgements:

Paraskevi Fragopoulou, [Vasilios Siris](#), and Georgios Sapountzis helped us shape our ideas; we deeply thank them.

© Copyright 2002-2003 by FORTH, Univ. of Crete, and IEEE:

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 FORTH or Univ. of Crete or IEEE copyright notice, the title of the publication and its date appear, and notice is given that copying is by permission of the Foundation for Research & Technology -- Hellas (FORTH) and/or the University of Crete and/or the IEEE. 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 Buffered Crossbar \(CICQ\) Switch Architecture at FORTH](#)

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