RETROSPECTIVE: Low-Latency Virtual-Channel Routers for On-Chip Networks

Robert Mullins*, Andrew West and Simon W. Moore*
*Department of Computer Science and Technology, University of Cambridge

I. Background

This paper was published at an interesting time for computer architecture just prior to the mainstream shift to multi-core processors. Prior to the work on on-chip interconnection networks our research group was exploring novel approaches to system timing and asynchronous circuit techniques. A collaboration with Prof. Krste Asanovic's group at MIT had been prompted by the Cambridge-MIT Initiative and started in January 2003. The goal was to develop new on-chip communication techniques. Asanovic's MIT group was working on the SCALE processor [11] at the time. This was a vector-thread architecture which was also designed to support the tiling of multiple SCALE processors in a CMP.

A number of very productive visits were organised between Cambridge and MIT and it was great to get to know and learn from the MIT team (including Krste, Ronny Krashinsky, Christopher Batten and many others). We also met with members of the RAW processor team at MIT, including Michael Taylor and David Wentzlaff. RAW [17] is an early tiled multi-core architecture, with a number of both statically and dynamically scheduled on-chip networks.

II. Evolution and later related work

The ISCA paper was actually the result of trying to understand the problem and establish the best possible base case before exploring implementations that exploited novel approaches to system-timing (e.g. asynchronous circuits and clocks that can be paused [14]). None of the Cambridge team had worked on any form of networking project before and came to the problem afresh. Initial work focused on Prof. Bill Dally's work on on-chip networks and the work of his then PhD student Li-Shuan Peh (now Prof. of Computer Science at the National University of Singapore).

We focused on the problem of low-latency on-chip communication using networks that exploited virtual-channels (VCs). Dally's VC flow control could be implemented simply as a three-stage pipeline. Dally and Peh had recently reduced this pipeline to two-stages, allowing VC and switch allocation to take place in parallel through the use of speculation [16].

Our contribution was to produce a single-cycle design. VC and switch allocation were performed concurrently with the transport of flits across the router's datapath. This was possible as scheduling decisions can often be performed ahead of time or in the lightly loaded case we could simply speculate that requests will be granted. Results showed that such an approach sacrificed only a few percent of performance over a perfect single-cycle design.

A. Validation in silicon

We refined our single-cycle router design and produced a 16-node test-chip, which we published at ASP-DAC in 2006 [15]. The designed also included a novel distributed clock generator (DCG) [6]. In 2016 the paper was awarded the ASP-DAC ten year retrospective most influential paper award.

B. Refinement and simplification

A few years after the ISCA publication, Rob realised an efficient VC router could be constructed using only a switch allocator. In this architecture, a virtual-channel was allocated when a waiting packet was granted access to an output port for the first time. To ensure performance does not suffer, switch requests from packets not yet assigned a VC are only made if the VC free pool at the required output port is not empty. It transpired that Amit Kumar in Li-Shuan Peh's group at Princeton had realised the same thing a little while before. This VC allocator design was included in a router design that Peh's team had developed with Intel [12]. Kumar also described an implementation with separate speculative and non-speculative allocators which they determined was better in terms of its critical path delay.

It is perhaps surprising that this simple and effective approach wasn't noted earlier by researchers or reviewers. We believe this approach supersedes ours, i.e. there is no performance degradation or other concerns.

III. Follow-on research at University of Cambridge

This ISCA paper was part of a journey into novel computer architectures. Our later work explored energy and performance optimization of on-chip networks [2], [3], flow-aware allocation [1], fault-tolerance [9], on-chip networks for FPGAs [7], [8] and low power photonic networks for distributed shared memory systems [18].

On-chip networks route packets not wires, but we questioned whether communication scaling patterns would still follow Rent's rule [5], [13] used to predict wire scaling for chips. For his Ph.D., Daniel Greenfield explored this question resulting in his thesis entitled *Communication Locality in Computation: Software, Chip Multiprocessors and Brains*, that

won the UK BCS Distinguished Dissertation Prize. Dr Greenfield discovered that Rent's rule also applied to scaling of on-chip networks and, surprisingly, it also predicts the scaling of interconnect in mammalian brains [4] and communication patterns in large software systems [10].

REFERENCES

[1] A. Banerjee and S. W. Moore, "Flow-aware allocation for on-chip networks," in *2009 3rd ACM/IEEE International Symposium on Networks-on-Chip*, 2009, pp. 183–192.

[2] A. Banerjee, R. Mullins, and S. Moore, "A power and energy exploration of network-on-chip architectures," in *First International Symposium on Networks-on-Chip (NOCS'07)*, 2007, pp. 163–172.

[3] A. Banerjee, P. T. Wolkotte, R. D. Mullins, S. W. Moore, and G. J. M. Smit, "An energy and performance exploration of network-on-chip architectures," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 17, no. 3, pp. 319–329, 2009.

[4] D. S. Bassett, D. L. Greenfield, A. Meyer-Lindenberg, D. R. Weinberger, S. W. Moore, and E. T. Bullmore, "Efficient physical embedding of topologically complex information processing networks in brains and computer circuits," *PLoS computational biology*, vol. 6, no. 4, p. e1000748, 2010.

[5] P. Christie and D. Stroobandt, "The interpretation and application of Rent's rule," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 8, no. 6, pp. 639–648, 2000.

[6] S. Fairbanks and S. Moore, "Self-timed circuitry for global clocking," in *Proc. Intl. Symp. on Asynchronous Circuits and Systems (ASYNC)*, 2005.

[7] R. Francis and S. Moore, "Exploring hard and soft networks-on-chip for fpgas," in *2008 International Conference on Field-Programmable Technology*, 2008, pp. 261–264.

[8] R. Francis, S. Moore, and R. Mullins, "A network of time-division multiplexed wiring for fpgas," in *Second ACM/IEEE International Symposium on Networks-on-Chip (nocs 2008)*, 2008, pp. 35–44.

[9] D. Greenfield, A. Banerjee, J.-G. Lee, and S. Moore, "Implications of rent's rule for noc design and its fault-tolerance," in *First International Symposium on Networks-on-Chip (NOCS'07)*, 2007, pp. 283–294.

[10] D. Greenfield and S. W. Moore, "Fractal communication in software data dependency graphs," in *Proceedings of the Twentieth Annual Symposium on Parallelism in Algorithms and Architectures*, ser. SPAA '08. New York, NY, USA: Association for Computing Machinery, 2008, p. 116–118. [Online]. Available: https://doi.org/10.1145/1378533.1378555

[11] R. Krashinsky, C. Batten, M. Hampton, S. Gerding, B. Pharris, J. Casper, and K. Asanovic, "The vector-thread architecture," in *Intl. Symp. on Computer Architecture (ISCA)*, 2004.

[12] A. Kumar, P. Kundu, A. P. Singh, L.-S. Peh, and N. K. Jha, "A 4.6Tbits/s 3.6GHz single-cycle NoC router with a novel switch allocator in 65nm CMOS," in *Intl. Conf. on Computer Design (ICCD)*, 2007.

[13] B. Landman and R. Russo, "On a pin versus block relationship for partitions of logic graphs," *IEEE Transactions on Computers*, vol. C-20, no. 12, pp. 1469–1479, 1971.

[14] R. Mullins and S. Moore, "Demystifying data-driven and pausible clocking schemes," in *Proc. Intl. Symp. on Asynchronous Circuits and Systems (ASYNC)*, 2007.

[15] R. Mullins, A. West, and S. Moore, "The design and implementation of a low-latency on-chip network," in *Proc. of Asia and South Pacific Design Automation Conference (ASP-DAC)*, 2006.

[16] L.-S. Peh and W. J. Dally, "A delay model and speculative architecture for pipelined routers," in *Intl. Symp. on High-Performance Computer Architecture (HPCA)*, 2001.

[17] M. Taylor, J. Psota, A. Saraf, N. Shnidman, V. Strumpen, M. Frank, S. Amarasinghe, A. Agarwal, W. Lee, J. Miller, D. Wentzlaff, I. Bratt, B. Greenwald, H. Hoffmann, P. Johnson, and J. Kim, "Evaluation of the Raw microprocessor: an exposed-wire-delay architecture for ILP and streams," in *Intl. Symp. on Computer Architecture (ISCA)*, 2004.

[18] P. Watts, N. Barrow-Williams, and S. Moore, "Requirements of low power photonic networks for distributed shared memory computers," in *2011 Optical Fiber Communication Conference and Exposition and the National Fiber Optic Engineers Conference*, 2011, pp. 1–3.