An Asynchronous NoC Router in a 14nm FinFET Library: Comparison to an Industrial Synchronous Counterpart

Weiwei Jiang*, Davide Bertozzi†, Gabriele Miorandi†, Steven M. Nowick*, Wayne Burleson‡ and Greg Sadowski‡

*Columbia University, New York, NY, USA {wjiang, nowick}@cs.columbia.edu
†University of Ferrara, Ferrara, Italy {davide.bertozzi, gabriele.miorandi}@unife.it
‡Advanced Micro Devices, Inc., Boxborough, MA, USA {wayne.burleson, greg.sadowski}@amd.com

***Abstract*—An asynchronous high-performance low-power 5-port network-on-chip (NoC) router is introduced. The proposed router integrates low-latency input buffers using a circular FIFO design, and a novel end-to-end credit-based virtual channel (VC) flow control for a replicated switch architecture. This asynchronous router is then compared to an AMD synchronous router, in a realistic advanced 14nm FinFET library. This is the first such comparison, to the best of our knowledge, using a real synchronous router baseline already fabricated in several commercial products. Initial post-synthesis pre-layout experiments show dominating results for the asynchronous router, when compared to the synchronous router. In particular, 55% less area and 28% latency improvement are observed for the asynchronous implementation. Also, 88% and 58% savings in idle and active power, respectively, are obtained.**

I. Introduction

Over the last decade, networks-on-chip have become a standard approach for on-chip communication. These networks typically use packet switching in a structured architecture, and inherently separate the computational elements from the communication infrastructure [3]. Recently, there has been increasing interest in building asynchronous NoCs, since they eliminate global clock management across a large network, and are therefore a natural match for NoC approaches [1][3][6].

Several recent commercial asynchronous and globally-asynchronous locally-synchronous (GALS) NoCs have been proposed: Intel's FM5000/6000 series Ethernet switch chips [6]; IBM's TrueNorth neuromorphic chip, modeling 1M neurons and 256M synapses with 4096 neurosynaptic cores [4]; STMicroelectronics' STHORM processor, an accelerator-based many-core GALS system [3]. These industrial examples exhibit flexible integration of heterogeneous components, as well as significant power and area benefits.

This paper provides the first comparison for an asynchronous router vs. an industrial synchronous baseline using an advanced technology library. While the library is state-of-art, the routers use simple structures and avoid advanced optimizations (lookahead/speculation), yet still achieve fairly high performance. Unlike other baselines for research purposes, the synchronous design is used in recent high-end AMD processors and graphic products, to handle system-level configuration and power/performance monitoring and control. The results are thus more persuasive and closer to reality. In addition, industrial tools are used for place-and-route (P&R) and design validation. These tools are modified from a standard synchronous design flow and therefore open real future opportunities for industrial asynchronous NoC designs. Also, the new asynchronous router contains a novel end-to-end credit-based VC control with potentially higher throughput.

II. Background and Overview of the Approach

The section reviews our foundational 5-port VC-less asynchronous router [1] on which we build. It also reviews our previous state-of-the-art credit-based VC control [5], and outlines our new optimized VC strategy.

This work was partially supported by NSF Grant CCF-1527796, and by the Italian Government through a 'Fondo Giovani' fellowship.

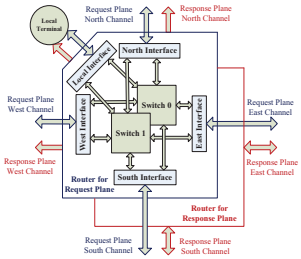

Fig. 1. Node structure for proposed asynchronous double-plane router

Our original router [1] explores a unique design point for asynchronous NoCs, using two-phase handshaking and bundled data encoding. This direction has recently shown promise by other groups: Imai and Yoneda [6] adopt this protocol, but limit its use to intra-switch, while using an inter-switch channel encoding with higher cost; the VC-less BAT-Hermes [2] includes realistic input buffers, but has expensive packet control and crossbar transmission. A two-phase protocol [6] is used for our proposed router, both intra- and inter-switch, which has only a single round-trip channel communication per transmission. Single-rail bundled data encoding [6] provides coding efficiency nearly identical to a synchronous design. In the new design, 5 Input Port Modules (IPMs) are connected through a crossbar to 5 Output Port Modules (OPMs). An IPM computes routing and propagates it to the designated OPM, while broadcasting it to all OPMs. An OPM identifies a valid request, and resolves arbitration between competing requests.

The new router also inherits a replicated-switch credit-based VC architecture [5]. As shown in Fig. 1, switches are replicated as many times as the number of VCs, e.g., Switches 0/1. (It also uses a double-plane structure; see Sec. III below.) VCs separate different traffic classes inside the router, which are mixed only on inter-router links. This structure outperforms a crossbar-sharing approach for asynchronous routers [5], although the latter is a typical approach for synchronous.

A new credit-based VC control is proposed. In [5], the credit count is decreased when a flit is sent out and is increased when the successor releases an input buffer slot. These two operations are mutually exclusive, and treated symmetrically at the same priority. A credit-increment operation can thus potentially block credit-decrement, and delay sending out a flit [5]. In contrast, the new approach only updates the credit when a flit is sent out. Credit-increment requests are queued and are only updated along with the next credit-decrement request. This 'lazy-update' scheme prevents unnecessary credit-increment updates and potentially increases the throughput.

III. Proposed Asynchronous Router Design

The router is designed for a 2D mesh with a double-plane NoC, the same structure as the synchronous network, which

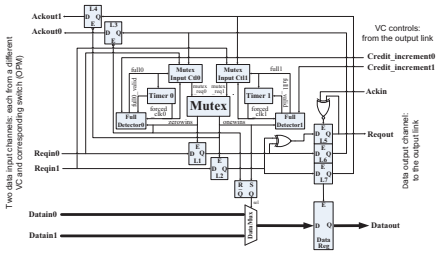

Fig. 2. Proposed VC control for an output channel interface

contains two uncorrelated and identical networks, as shown in Fig. 1. The request plane routes request packets and the response plane delivers responses. In each router, the switch is identical to the original switch [1] but with an extra circular input buffer added. The packets are routed to the appropriate switch (in the same plane), based on its statically-assigned VC.

Input buffer. An optimized circular FIFO implementation, presented in [1], is now integrated into the proposed design.

VC flow control. Fig. 2 shows the new VC flow control. It takes input streams from 2 VCs, performs flit-level arbitration and merges them to a single output stream. The *Full Detector* updates the credit every time a flit is sent out. The update considers all queued credit-increment requests as well as the current credit-decrement. If there is no credit, the VC is blocked; *Timer* is activated, which constantly checks at a fixed rate if any credits are released; if so, blocking is released.

IV. Design Flow and Tools

Design validation tool. Detailed gate-level functional validation is performed for both pre- and post-layout designs for a single double-plane router node. A synchronous industrial tool is used, with a new wrapper to synchronize the I/O data to an external clock. Hence, the asynchronous router with wrapper can simply be plugged into the existing tool, and standard benchmarks used, as in a synchronous testing flow.

Design flow and P&R tool. The design was first manually synthesized by mapping each gate to a real library. Asynchronous one-sided timing constraints are satisfied by manually adding proper inverter-chain delays. Manual mapping prevents logic optimization, which can potentially create control glitches. Research solutions for asynchronous logic synthesis automation have been proposed [1], which are not included due to the extensive effort required to re-instrument the stable industrial flow. However, it is expected that no serious obstacles appear to their inclusion in the future.

The P&R flow uses a standard automated synchronous approach, without logic optimization. The final asynchronous layout is shown in Fig. 3, which was used for design validation.

Although post-layout results could not be reported at this time, due to the use of advanced commercial technology, this flow demonstrates the viability of incorporating asynchronous physical design into a leading industrial environment. Furthermore, the strong initial asynchronous results are highly encouraging, and expected to contribute to industrial motivation to invest in asynchronous CAD tool development.

V. Experimental Results

Comparisons are performed for a single pre-layout router node in terms of area, latency and power. The synchronous baseline is a typical 3-cycle router, with fine-grain clock gating. It also has some additional functionality for error detection and router configuration; these contribute only 1-4% area and power increase, with negligible performance impact. All results are presented in relative numbers only, due to

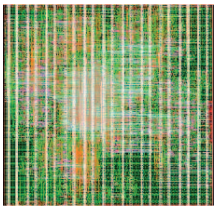

Fig. 3. Actual layout for the proposed asynchronous router

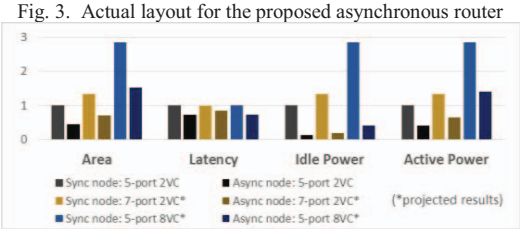

Fig. 4. Asynchronous and synchronous router comparison

confidentiality reasons. Based on industrial experience, it is expected that pre-layout comparisons and post-layout comparisons will be similar for such a small router design.

For the basic comparison, both routers have 2 VCs, each with buffer depth of 7. Each is synthesized using a low-power industrial 14nm library (0.65V, TT corner, 273K). The synchronous router is synthesized targeting a 1 GHz clock rate, based on the performance requirements of several high-end AMD products, using a standard automated flow, while the asynchronous router is synthesized manually, as indicated. Evenly distributed random traffic is sent from all input ports to output ports, with a random packet size between 2 and 6.

In Fig. 4, the left two bars of each group show outcomes for the basic comparison. The asynchronous router dominates the results: 55% lower area and 28% latency improvement, with 88% and 58% savings in idle and active power, respectively.

Fig. 4 also presents some estimated results for (i) a 7-port router with 2 VCs, and (ii) a 5-port router with 8 VCs. The former is important for 3D stacking, and the latter represents a more realistic VC configuration. For both synchronous and asynchronous routers, area and power costs noticeably increase in (i) and (ii), due to higher radix or more VCs, while latency is largely unchanged. However, relative asynchronous area and power benefits are largely maintained, though latency improvements are reduced for the 7-port configuration.

VI. Conclusions

The paper presents the first comparison of an asynchronous vs. commercial synchronous NoC router in an advanced technology. The new design uses several industrial tools for P&R and design validation. A novel end-to-end credit-based VC control is also included. Results show the asynchronous router obtains significant benefits in area, latency and power.

References

[1] A. Ghiribaldi, D. Bertozzi, and S.M. Nowick. A transition-signaling bundled data NoC switch architecture for cost-effective GALS multicore systems. In *Proc. of DATE Conf.*, pp. 332-227, 2013.

[2] M. Gibiluka, M.T. Moreira, F.G. Moraes and N.L.V. Calazans. BAT-Hermes: a transition-signaling bundled-data NoC router. In *Proc. of Latin American Symposium on Circuits & Systems,* pp. 1-4, 2015.

[3] D. Melpignano, L. Benini, E. Flamand and F. Clermidy *et al.* Platform 2012, a many-core computing accelerator for embedded SoCs: performance evaluation of visual analytics applications. In *Proc. of DAC Conf.,* pp. 1137-1142, 2012.

[4] P. Merolla *et al.* A million spiking-neuron integrated circuit with a scalable communication network and interface. In *Science,* 345(6197):668-673, 2014.

[5] G. Miorandi, A. Ghiribaldi, S.M. Nowick, and D. Bertozzi. Crossbar replication vs. sharing for virtual channel flow control in asynchronous NoCs: a comparative study. In *IFIP/IEEE VLSI-SoC Conf.*, pp. 1-6, 2014.

[6] S.M. Nowick and M. Singh. Asynchronous design – part I: overview and recent advances. In *IEEE Design and Test*, 22(3):5-18, 2015.