

SmartFork: Partitioned Multicast Allocation and Switching in Network-on-Chip Routers

Dimitrios Konstantinou*, Chrysostomos Nicopoulos†, Junghee Lee‡, Georgios Ch. Sirakoulis*, Giorgos Dimitrakopoulos*

*Electrical and Computer Engineering, Democritus University of Thrace, Xanthi, Greece

†Electrical and Computer Engineering, University of Cyprus, Nicosia, Cyprus

‡School of Cyber Security, Korea University, Seoul, Korea

Abstract—Multicast on-chip communication is encountered in various cache-coherence protocols targeting multi-core processors, and its pervasiveness is increasing due to the proliferation of machine learning accelerators. In-network handling of multicast traffic imposes additional switching-level restrictions to guarantee deadlock freedom, while it stresses the allocation efficiency of Network-on-Chip (NoC) routers. In this work, we propose a novel NoC router microarchitecture, called SmartFork, which employs a versatile and cost-efficient multicast packet replication scheme that allows the design of high-throughput and low-cost NoCs. The design is adapted to the average branch splitting observed in real-world multicast routing algorithms. Compared to state-of-the-art NoC multicast approaches, SmartFork is demonstrated to yield higher performance in terms of latency and throughput, while still offering a cost-effective implementation.

I. INTRODUCTION

The increase in the number of processing elements in modern microprocessors and Systems-on-Chip (SoC) has accentuated the criticality of the on-chip interconnect, in terms of both performance and functional correctness. Networks-on-Chip (NoC) have been established as the enabling communication fabrics that can sustain the many-core era. Multicast (and often broadcast) communication is widespread in some of the most popular cache-coherence protocols targeting multi-core processors. The multicast intensity has been shown to increase with the number of on-chip cores, thereby underscoring its performance impact on system scalability. Furthermore, multicast traffic is also widespread in the increasingly prevalent hardware accelerators targeting artificial intelligence [1].

A. Multicast-enabled NoCs: The current state-of-the-art

In-network multicast support can take various forms. A naive and low-performing approach is to inject multiple unicast clones of the packet, with each one sent to a distinct recipient (i.e., *unicast-based*). To increasing performance – while keeping a minimal multicast-packet footprint – one may employ *path-based* multicast routing [2], [3], [4]. Only one multicast packet is injected, which sequentially visits all recipients. The packet is delivered to a single recipient at a time, allowing for at most two multicast branches, with one of them always sinking into the local ejection port.

To further improve performance, *tree-based* multicasting relaxes the replication degree, allowing for branching to an arbitrary number of output ports within each router [5], [6], [7], [8], [9]. This extra flexibility enables each recipient to be served independently, and, thus, more quickly, as opposed to creating sequential dependencies among recipients.

Nevertheless, the increased branching flexibility in tree-based multicast algorithms is precisely the reason why deadlocks may, in fact, arise at the multicast-replication level, i.e., as a result of dependencies among the various multicast branches. Note that such switching-level deadlocks arise even if the routing algorithm is deadlock-free. To tackle this issue, Virtual Cut-Through (VCT) switching [10], [11], [12] is employed. Alternatively, low-performance circuit-switching approaches [13], [14], or costly deadlock recovery schemes [15], have also been presented.

Multicast packets in NoCs are generally short, as they typically carry only control information [16], or single-word data; e.g., a cacheline invalidation message. Hence, multicast messages in NoCs could most likely fit within a single-flit packet. Single-flit packets make the routing of each multicast branch independent, and their switching in each router is, *by construction*, deadlock-free. This attribute has been exploited in [6], [9] to build multicast NoCs. In a similar vein, the work in [17] transforms all multicast packets into independent single-flit multicast packets.

Single-flit multicast allocation and switching inside each router can be performed in two ways. One approach is to treat multicast branch replication as a set of serially-executed unicast transmissions; i.e., sending a flit to multiple output ports in the same cycle is prohibited [11]. This branch serialization yields a low-cost multicast solution, but limits the flit replication rate to one output port per cycle. Alternatively, each multicast packet can be replicated in parallel to all required output branches in each router [9]. This parallel replication is readily supported by the crossbar of each NoC router. However, the increase of requests from multiple input virtual channels to multiple output virtual channels stresses the separable allocators used in state-of-the-art NoC routers [18], thereby increasing Head-of-Line (HoL) blocking.

To address the inefficiency of separable allocation when dealing with mixed multicast and unicast traffic, the use of input buffers with as many independent read ports as the number of output ports in the router (aka *buffer speedup*) has been proposed [10], [12]. This approach increases throughput, but with prohibitive hardware cost – especially with respect to local wiring congestion.

B. How much parallel replication is adequate?

The analysis above indicates that an ideal multicast-enabled NoC design would reap all the benefits of parallel branch replication (as used by state-of-the-art NoC routers), without resorting to full-fledged buffer speedup. Before embarking on any micro-architectural optimizations targeting such goal, it

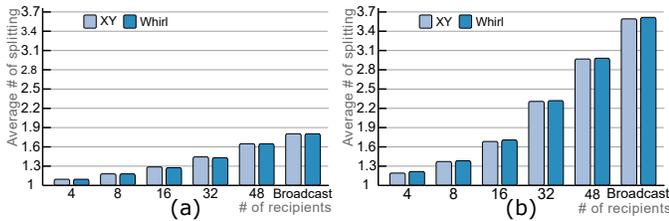


Fig. 1. The average number of branch splits per router required by the XY and Whirl [9] multicast routing algorithms for different number of multicast destinations in (a) an 8×8 2D mesh, and (b) a 64-node 2D mesh with high-radix routers (8-port routers with a concentration factor of 4).

is important to investigate how multicast routing algorithms behave, in terms of the degree of observed multicast branch splitting per router.

Parallel branch replication per router – with or without buffer speedup – assumes that each input port can be connected in parallel to all output ports. For example, in a typical 5-port router employed in 2D mesh NoCs, each input would be connected to 5 outputs. Nevertheless, the key question is whether real-world multicast routing algorithms exhibit such high levels (5-way) of branch splitting in each router. In fact, the amount of multicast branch splitting observed in 2D meshes (either low-radix, or high-radix) under the most established and widely used multicast routing algorithms is much lower than 5. This argument is experimentally verified in Fig. 1(a) for an 8×8 2D mesh under two different state-of-the-art multicast routing algorithms (XY and Whirl [9]), and under different multicast intensities. Note that Whirl’s branch trees constitute a superset of the trees produced by several other multicast routing algorithms [9]. A similar conclusion regarding branch splitting can be drawn for higher-radix topologies, as shown in Fig. 1(b), which assumes a router concentration degree of 4 in a 64-node NoC, i.e., each router has 8 output ports. These low levels of intra-router multicast branch splitting are orthogonal to the NoC traffic characteristics and the application workloads running on the multicore system. *The degree of multicast branch splitting is inherently an attribute of the routing algorithm itself.*

C. Contributions

Driven by the observation that the degree of branch replication per router, as dictated by the multicast routing algorithm, is much lower than the number of input/output ports in a router, we hereby propose *SmartFork*, a scalable NoC micro-architecture that supports multicasting and is characterized by the following novel features:

(1) The output ports are partitioned into groups; intra-group ports are serviced serially, while inter-group ports are serviced in parallel. Therefore, SmartFork can be used to target any desired point on the multicast performance spectrum.

(2) Each output partition is driven by a separate input-buffer read port, thus mitigating allocation inefficiencies without introducing prohibitive hardware cost.

(3) Virtual Channels (VC) are inherently supported without any restrictions on which VCs can support multicasting; any VC within the router can support multicast transfers.

The efficacy and efficiency of SmartFork are corroborated through extensive cycle-accurate network simulations and detailed hardware analysis of synthesized and placed-and-routed designs using commercial 45 nm standard-cell libraries.

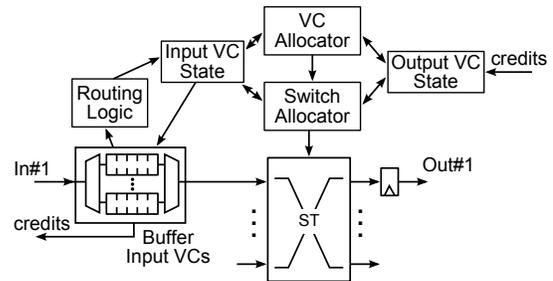


Fig. 2. The micro-architecture of a typical virtual-channel-based NoC router.

II. SMARTFORK: PARTITIONED MULTICASTING

The NoC routers facilitate the transfer of packets between communicating (source/destination) nodes. Packets are typically split into smaller flow-control units, aka *flits*. The flits traverse the network in a hop-by-hop fashion through all the routers encountered on the path from the source to the destination. The micro-architecture of a typical VC-based NoC router [18] is depicted in Fig. 2.

Each router has N input and N output ports. In a 2D mesh NoC topology, $N = 5$; one input/output port for each cardinal direction, and a local injection/ejection port. Each input port has V virtual channels, with each channel being served by a different FIFO input buffer. All incoming packets are written into these input buffers, and they pass through a series of operational stages before they can exit the router. Routing Computation (RC) logic determines the output port of each packet. The Virtual-channel Allocation (VA) stage maps input VCs to output VCs. The Switch Allocation (SA) stage – that is typically split in two stages, SA1 and SA2 – arbitrates amongst all input VCs requesting access to the crossbar, and declares one winning flit for each input and output port. The SA winners are then able to traverse the crossbar (Switch Traversal, ST), and are placed on their respective output links.

The basic organization shown in Fig. 2 can, in fact, be used to also support *multicast* transmissions, as explained in [9]. This is achieved by allowing the flits of each input VC to simultaneously send requests to *multiple* output ports. For instance, if a multicast packet must branch out to three output ports of the router (as decided by the employed multicast routing algorithm), then the packet is allowed to send concurrent requests to all three output ports. While said approach is effective in handling multicast communication and is cost-efficient, it suffers from an often-debilitating affliction: increased susceptibility to HoL blocking. Specifically, if any of the multiple requested output ports is/are unavailable, the flit at the head of the input buffer must remain there until *all required output ports* are, eventually, served. Only then can the flit be popped from the input buffer. Consequently, all the other flits occupying the same input buffer are HoL-blocked.

A. Router organization

SmartFork is founded on *two basic properties* that enable high multicasting efficiency with minimal changes to the baseline VC-based NoC router microarchitecture. The first property places an upper limit P on the number of multicast branches that can be served in parallel in any given cycle. The value of P aims to reflect the average branching degree per router observed in established multicast routing algorithms. To achieve this in a cost-efficient manner, the N output ports in a

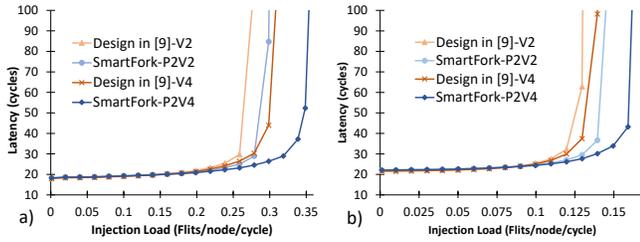


Fig. 4. Network performance results for routers with 2 and 4 VCs under synthetic uniform-random traffic with (a) 5%, and (b) 30% multicast traffic intensity in an 8×8 2D mesh NoC.

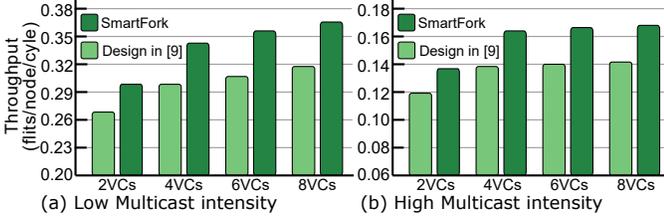


Fig. 5. Saturation throughput comparison of the router architectures utilizing up-to 8 VCs for (a) 5% and (b) 30% multicast traffic intensity in an 8×8 2D mesh NoC.

network nodes, which is in line with what has been observed in real applications [23].

The results pertaining to the synthetic uniform-random traffic are depicted in Figs. 4(a) and (b), for 5% and 30% multicast traffic intensities, respectively. Compared to the state-of-the-art [9], SmartFork with 2 VCs per input port yields a throughput increase of 11% and 13% when 5% and 30% multicast traffic intensity is applied, respectively. When 4 VCs are used, the reaped gains are amplified to 15% and 18%, respectively. To further investigate the designs' scalability, Fig. 5 depicts the saturation throughput achieved as the VCs per input port increase. The saturation throughput improves with increasing VC numbers, albeit with diminishing returns. Nevertheless, SmartFork achieves up to 20% throughput improvement over the state-of-the-art [9] when 8 VCs are used. The results of Figs. 4 and 5 highlight the scalability of SmartFork with the number of VCs: as the VCs increase, the design in [9] experiences increasing HoL blocking. On the contrary, SmartFork effectively mitigates HoL blocking and reaps substantial throughput improvements.

For the experiments with *cache-coherence* traffic, the model developed in-house generates traffic that mimics the behavior of (1) HyperTransport [24] (a directory protocol), and (2) Token Coherence [25] (a snoopy protocol), in a 64-node CMP. A significant portion of the injected traffic under the HyperTransport and Token Coherence protocols is broadcast traffic: 6% and 15%, respectively. Under cache-coherence traffic, the SmartFork design with 2 VCs achieves throughput improvements of 11% and 12%, while with 4 VCs the improvements are 15% and 19% for HyperTransport and Token Coherence traffic, respectively, as shown in Figs. 6(a) and (b).

The four 5-port NoC routers under comparison were synthesized using a commercial low-power 45 nm standard-cell library under worst-case conditions (0.8 V, 125 °C), and placed-and-routed using the Cadence digital implementation flow. In all NoC configurations, the flit width was set to 128 bits. The obtained results are summarized in Table I.

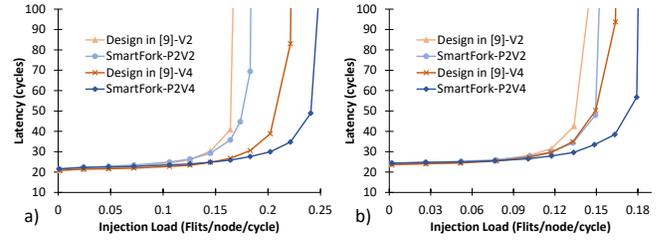


Fig. 6. Network performance results for routers with 2 and 4 VCs under traffic closely resembling (a) HyperTransport, and (b) Token Coherence traffic in an 8×8 2D mesh NoC.

To enable fair and meaningful comparisons with respect to area/power/energy, all four designs operate at 1 GHz, which is close to their maximum achievable clock frequency.

TABLE I
HARDWARE IMPLEMENTATION RESULTS OF THE ARCHITECTURES UNDER INVESTIGATION AT 45 NM TECHNOLOGY AND 0.8 V.

Design @ 1 GHz	VCs	Area (μm^2)	Power (mW)	Thru-put%Gain / Area%Incr.	Thru-put%Gain / Power%Incr.
Design in [9]	2	57223	5.36	1.39	3.09
SmartFork	2	62370	5.58		
Design in [9]	4	118393	10.10	1.73	12.37
SmartFork	4	130442	10.25		

As shown in Table I, SmartFork occupies slightly more area (around 9-10%), as compared to the state-of-the-art [9]. However, SmartFork achieves much higher throughput, and, due to its efficient micro-architecture, it incurs a near-negligible power consumption overhead. In fact, SmartFork's power consumption is very similar to that of the architecture in [9].

To evaluate the proposed SmartFork design's area and power efficiency, we utilize two metrics that are derivatives of the *Kill Rule* [26]. Said rule states that the percentage gain in performance (in our case, throughput) should outweigh the percentage increase in hardware cost (i.e., in area and power). Based on this reasoning, the two metrics employed are: (a) Throughput percentage gain over area percentage increase, and (b) Throughput percentage gain over power percentage increase. The obtained values for these metrics are shown in the last two columns in Table I, and they refer to the two SmartFork variants. Obviously, *all four values are larger than 1*, thereby indicating *higher throughput gain than the hardware cost paid*. For example, in the case of SmartFork with 4 VCs, every 1% increase in power consumption results in a 12.37% increase in reaped throughput. Overall, the two efficiency metrics clearly indicate that SmartFork is incredibly area- and power-efficient in increasing the achieved throughput.

IV. CONCLUSIONS

The increasing prevalence of multicast traffic in NoCs highlights the imperative need to provide scalable multicast support in future systems. In this work, we present the novel VC-based and multicast-enabled SmartFork NoC router architecture. SmartFork relies on a flexible and cost-efficient packet replication mechanism that can yield implementations targeting any desired point on the multicast performance spectrum. A specific variant of SmartFork – adapted to the average branch splitting observed under well-known multicast routing algorithms – is demonstrated to yield higher throughput than the current state-of-the-art architecture, while also achieving very high area and power efficiency.

REFERENCES

- [1] Arteris, "Arteris IP FlexNoC AI package," 2018.
- [2] X. Lin and L. M. Ni, "Deadlock-free multicast wormhole routing in multicomputer networks," *Comp. Archit. News*, pp. 116–125, 1991.
- [3] R. V. Boppana, S. Chalasani, and C. S. Raghavendra, "On multicast wormhole routing in multicomputer networks," in *IEEE Symp. on Par. and Distr. Processing*, Oct 1994, pp. 722–729.
- [4] M. Ebrahimi, M. Daneshlab, M. H. Neishaburi, S. Mohammadi, A. Afzali-Kusha, J. Plosila, and H. Tenhunen, "An efficient dynamic multicast routing protocol for distributing traffic in nocs," in *Design Automation and Test in Europe (DATE)*, April 2009, pp. 1064–1069.
- [5] N. E. Jerger, L.-S. Peh, and M. Lipasti, "Virtual circuit tree multicasting: A case for on-chip hardware multicast support," *Comput. Archit. News*, pp. 229–240, 2008.
- [6] S. Ma, N. E. Jerger, and Z. Wang, "Supporting efficient collective communication in nocs," in *IEEE International Symposium on High-Performance Comp Architecture (HPCA)*, 2012.
- [7] S. Rodrigo, J. Flich, J. Duato, and M. Hummel, "Efficient unicast and multicast support for cmps," in *Intern. Symp. on Microarchitecture (MICRO)*, Nov 2008, pp. 364–375.
- [8] L. Wang, Y. Jin, H. Kim, and E. J. Kim, "Recursive partitioning multicast: A bandwidth-efficient routing for networks-on-chip," in *Int. Symp. on Networks-on-Chip (NoCS)*, May 2009, pp. 64–73.
- [9] T. Krishna, L.-S. Peh, B. M. Beckmann, and S. K. Reinhardt, "Towards the ideal on-chip fabric for 1-to-many and many-to-1 communication," in *Int. Symp. on Microarchitecture (MICRO)*, 2011, pp. 71–82.
- [10] R. Sivaram, D. Panda, and C. Stunkel, "Multicasting in irregular networks with cut-through switches using tree-based multidestination worms," in *Parallel Computer Routing and Communication*, 1997, pp. 39–52.
- [11] W. Hu, Z. Lu, A. Jantsch, and H. Liu, "Power-efficient tree-based multicast support for networks-on-chip," in *ASP-DAC*, Jan 2011, pp. 363–368.
- [12] K. Bhardwaj, W. Jiang, and S. M. Nowick, "Achieving lightweight multicast in asynchronous nocs using a continuous-time multi-way read buffer," in *Int. Symp. on Networks-on-Chip (NoCS)*, 2017.
- [13] Z. Lu, B. Yin, and A. Jantsch, "Connection-oriented multicasting in wormhole-switched networks on chip," in *ISVLSI*, March 2006.
- [14] R. A. Stefan, A. Molnos, and K. Goossens, "daelite: A tdm noc supporting qos, multicast, and fast connection set-up," *IEEE Trans. on Computers*, pp. 583–594, March 2014.
- [15] M. P. Malumbres, J. Duato, and J. Torrellas, "An efficient implementation of tree-based multicast routing for distributed shared-memory multiprocessors," in *IEEE Symp. on Par. and Dist. Processing*, Oct 1996.
- [16] S. Ma, N. E. Jerger, and Z. Wang, "Whole packet forwarding: Efficient design of fully adaptive routing algorithms for networks-on-chip," in *IEEE International Symposium on High-Performance Comp Architecture (HPCA)*, Feb 2012, pp. 1–12.
- [17] F. A. Samman, T. Hollstein, and M. Glesner, "Adaptive and deadlock-free tree-based multicast routing for networks-on-chip," *IEEE Trans. on VLSI Systems*, pp. 1067–1080, July 2010.
- [18] G. Dimitrakopoulos, A. Psarras, and I. Seitanidis, *Microarchitecture of network-on-chip routers: A designer's perspective*. Springer, 2015.
- [19] I. Seitanidis, A. Psarras, E. Kalligeros, C. Nicopoulos, and G. Dimitrakopoulos, "ElastiNoC: A self-testable distributed vc-based network-on-chip architecture," in *International Symposium on Networks-on-Chip (NOCS)*, 2014, pp. 135–142.
- [20] A. Psarras, I. Seitanidis, C. Nicopoulos, and G. Dimitrakopoulos, "Short-Path: A network-on-chip router with fine-grained pipeline bypassing," *IEEE Transactions on Computers*, vol. 65, no. 10, pp. 3136–3147, 2016.
- [21] G. Dimitrakopoulos, N. Chrysos, and C. Galanopoulos, "Fast arbiters for on-chip network switches," in *IEEE Intern. Conf. on Computer Design (ICCD)*, 2008, pp. 664–670.
- [22] G. Dimitrakopoulos, E. Kalligeros, and K. Galanopoulos, "Merged switch allocation and traversal in network-on-chip switches," *IEEE Transactions on Computers*, vol. 62, no. 10, pp. 2001–2012, 2013.
- [23] S. Abadal, R. Martínez, J. Solé-Pareta, E. Alarcón, and A. Cabellos-Aparicio, "Characterization and modeling of multicast communication in cache-coherent manycore processors," *Computers and Electrical Engineering*, 2016.
- [24] P. Conway and B. Hughes, "The amd opteron northbridge architecture," *IEEE Micro*, vol. 27, 2007.
- [25] M. M. K. Martin, M. D. Hill, and D. A. Wood, "Token coherence: decoupling performance and correctness," in *International Symposium on Computer Architecture (ISCA)*, June 2003.
- [26] A. Agarwal and M. Levy, "The kill rule for multicore," in *Design Automation Conference (DAC)*, 2007, pp. 750–753.