The International Conference on Research Perspectives : IoT in Hybrid Grid Integrated Renewable Energy Sources
In association with International Journal of Scientific Research in Science, Engineering and Technology
Print ISSN: 2395-1990 | Online ISSN : 2394-4099 (www.ijsrset.com)

Network–On–Chip Router Microarchitecture for Future Communication : A Comprehensive Review

Ramanamma Parepalli[1], Dr. B Mohan Kumar Naik[2]

[1]Assistant Professor, Electronics and Communication Department, New Horizon college of Engineering , Bangalore, Karnataka, India

[2]Professor, Electronics and Communication Department, New Horizon college of Engineering , Bangalore, Karnataka, India

ABSTRACT

Network-on-Chip (NoC) is fast emerging as an on-chip communication alternative for many-core System-on-Chips (SoCs). NoC architecture is a preferable communication backbone for today's multiprocessor platforms. NoCs utilize routers at each node to direct traffic. However, designing a high performance, low latency NoC with low area overhead has remained a challenge. Conventional NoC router micro-architecture has main drawbacks in terms of circuit complexity, high critical path delay, resource utilization, timing, and power efficiency. The growing reliance on intellectual properties exposes SoCs to many security vulnerabilities and is raising more and more concerns. At the same time, with the quick increase in chip density and deep scaling of feature size, current billion-transistor chip designs introduce more challenges to manufacturing fault-free chips. The research presented in this paper has investigated these issues in detail and wants to develop a low latency, low-power and high-performance NoC router architecture that is applicable to a wide range of FPGA families.

Keywords - NoC router, VC allocator, Switch allocator, Switch traversal.

I. INTRODUCTION

As the quantity of cores on a single chip keep increasing, the network-on-chips (NoCs) technology is getting to be key to interconnect these cores. . NoC has turned out to be an essential piece of chip multiprocessors (CMPs) and interconnect hundreds even thousand cores [1]. Network-on-chip (NOC) solves the lack of scalability issue in bus-based interconnection system-on-chip (SOC) [2].Network-on-Chip (NoC) gives an adaptable and extensible inter-core-communication infrastructure for many-core system-on-chips. NoCs utilize routers at each node to direct traffic A typical, baseline router pipeline consists of two stages [3]. However, because of numerous numbers of routers a packet needs to navigate between a source and destination cores, as well as each individual router buffering, NoC-based frameworks can suffer from high inter-core communication latency. Reducing NoC communication latency while maintaining good throughput, a router needs to perform several stages such as route computation, VC allocation, and switch allocation in parallel. However, designing a low latency NoC router is still a challenge for on-chip systems [4].

Current NoC routers apply several virtual channels (VCs) on a solitary physical channel, for multiple

proposes for example, expanding system throughput, avoiding deadlock in fully adaptive routing [5], isolating resources for different message classes to prevent application-level deadlock [6], and enhancing Quality-of-service (QoS) by generating virtual networks [7]. VC makes the router architecture to become more complex that requires additional VC allocation stage to the existing router pipeline stages. In order to reduce on-chip memory usage, worm-hole flow control algorithm is widely applied in NoC routers [8-10]. Due to limited area and power budgets in a SoC, wormhole flow control is commonly used in NoC routers. Wormhole lowers the buffer requirement by storing different flits of the same packet in several routers along the path. Apart from that, NoC routers typically employ several VCs on a single physical channel as it offers several benefits such as increasing throughput by functioning as escape channels for active packets to bypass head-of-line (HoL) blocking, preventing deadlock condition in both network-level and protocol level as well as generating VNs to support QoS for different applications. A worm hole router divides a packet into several smaller flow control digits and allows flits to be buffered in a flit serial fashion order through several routers along the path. To provide low latency, there have been significant efforts on the design of routers [11], and network topologies. However, due to the stringent power and area budgets in a chip, simple routers and network topologies are more desirable.

In order to reduce the power utilization and increase the performance, various optimizations in the area of on-chip network design have been proposed. Different methods have been proposed for various on-chip network problems for example, layout designs, router microarchitecture design, mapping methods, and switching and flow control mechanism issues. Considering the aforementioned optimizations, flow control and buffering issues play a crucial role in power reduction [12].

II. LITERATURE REVIEW

RoB-Router: A Reorder Buffer Enabled LowLatency Network-on-Chip Router, IEEE Transactions on Parallel and Distributed Systems,2018 by Cunlu Li et al. [13] have developed a method to schedule packets in input buffers utilizing reorder buffer (RoB) techniques. The virtual channels VCs was designed as RoBsto allow packets located not at the head of a VC to be allocated before the head packets. RoBs reduce the conflicts in switch allocation and mitigate the HoL blocking and thus improve the NoC performance. However, it is hard to reorder all the units in a VC due to circuit complexity and power overhead. Then the RoB-Router was proposed, which leverages elastic RoBs in VCs to only allow a part of a VC to act as RoB. RoB-Router automatically determines the length of RoB in a VC based on the number of buffered flits. The design minimizes the resource while achieving excellent efficiency. Furthermore, two independent methods was proposed to improve the performance of RoB-Router. One was to optimize the packet order in input buffers by redesigning VC allocation strategy. The other combines RoB-Router with current most efficient switch allocator TS-Router. The method achieves 46% and 15.7% performance improvement in packet latency under synthetic traffic and traces from PARSEC than TS-Router, and the cost of energy and area was moderate.

ATAR: An Adaptive Thermal-Aware Routing Algorithm for 3-D Network-on-Chip Systems, IEEE Transactions on Components, Packaging and Manufacturing Technology, 2018 by Dash et al. [14] have developed an adaptive thermal-aware routing (ATAR) algorithm to distribute the traffic more uniformly across the chip for uniform thermal distribution in the chip. ATAR makes the decision based on the weighted sum approach by taking input parameters such as path length, next router

temperature, next router queue length, and link workload into account. Distributed ATAR units are coupled using a dynamic network to regulate the routing workload and results in minimization of thermal hotspots. In ATAR, cost computation and direction selection take place in accordance with a deadlock-free turn model. ATAR performs better in terms of thermal management and average packet delay.

Multicast-Aware High-Performance Wireless Network-on-Chip Architectures", IEEE Transactions on Very Large Scale Integration (VLSI) systems, 2017 by Duraisamy et al. [15] have presented a multicast-aware WiNoC architecture which can efficiently handle multicast-heavy cache coherence communications. Incorporated with a congestion-aware multicast routing and NC, WiNoC eliminates the initial and intermediate queuing latencies seen in conventional wire-line mesh NoCs. Moreover, using wireless shortcuts, the WiNoC achieves significant reductions in network latencies leading to improved system performances. The multicast aware WiNoC achieves an average of 47% reduction in message latency compared with the XY-tree-based multicast-aware mesh NoC.

OrthoNoC: A Broadcast-Oriented Dual-Plane Wireless Network-on-Chip Architecture", **IEEE Transactions on Parallel and Distributed Systems**, 2018 by SergiAbadal et al. [16] have presented ORTHONOC, a hybrid wired-wireless architecture composed of two independent network planes driven by a hybrid controller that can be agnostic or aware of the architecture. With the architecture-agnostic approach, ORTHONOC achieves significant speedups over other hybrid NoCs by offloading the wired plane from traffic for which it was inefficient. The method achieves 30 percent latency improvement, 25 percent throughput improvement, and higher energy efficiency with a similar number of wireless interfaces than other wired-wireless designs. With the architecture-aware approach, ORTHONOC's consistent order of delivery enables the design of faster and simpler multiprocessor architectures.

Randomly prioritized buffer-less routing architecture for 3D Network on Chip, Computers & Electrical Engineering, 2017 by Karthikeyan et al. [17] have presented a 3D lottery routing algorithm which was based on arbitral mechanism like randomly prioritized buffer. Communication among the IPs in NoC can be customized by users through the lottery router. The lottery routing algorithm distinguishes the different priorities of the input port and makes sure that it responses to the higher priority port. The efficient hardware implementation of 3D NoC was proposed using Xilinx Spartan 3E FPGA, the architecture consumes 1644 slices out of 4656 slices and operates at the maximum frequency of about 103.602MHz. The power consumption of 3D NoC was reduced by 9% compared to a single layer.

An Efficient Network-on-Chip Router for Dataflow Architecture", Journal of Computer Science and Technology, 2017 by Shen et al. [18] have proposed an efficient NoC router for dataflow architecture. The router supports multiple destination; thus, it can transfer data with multiple destinations in a single transfer. Moreover, the router adopts output buffer to maximize throughput and adopts non-flit packets to minimize transfer delay. The method improves the performance of dataflow architecture by 3.6 x over a state-of-the-art router.

A highly efficient dynamic router for application-oriented network on chip, The Journal of Supercomputing, 2018 by Su et al. [19] have proposed an effective router architecture including intra-port and inter-port allocation mechanism to improve network performance. By modifying the virtual channel unit of NoC's routers, the new router architecture can improve the buffer utilization

without affecting network performance. By using the idea of virtual output queue (VOQ), the router can solve the problem of head of line blocking. It can also balance the traffic load flexibly between different ports on the application-oriented NoC.

ProNoC: A Low Latency Network-on-Chip based Many-Core System-on-Chip Prototyping Platform, Microprocessors and Microsystems, October 2017 by Monemi et al. [20] have proposed a prototype NoC (ProNoC) method, Prototype NoC functionally represent an actual low latency ASIC-style NoC on field-programmable gate arrays (FPGA) platform. The method was developed in Register Transfer Level (RTL) and can support different features of a modern NoC such as VC, VN, different routing algorithms, non-atomic VC relocation, LRC, different VC and SW combined-allocator types as well as single clock cycle latency acceleration on packets traversing long distances. The proposed router was resource optimized to satisfy the limited hardware area of FPGA device while running at acceptable operating frequency. Pro NoC was also equipped with an EDA tool aided with GUI to facilitate NoC emulation and automated system integration of a complete NoC-based MC SoC in RTL level.

Low Latency Network-on-Chip Router Micro-architecture Using Request Masking Technique by Monemi et al. [21] have presented a two-clock-cycle latency router micro-architecture with parallel VC and switch allocator. The architecture eliminates the need of setting higher priority to any IVC requests. The NoC router architecture, any request which has been granted service by the switch allocator is able to pass a flit to the output port successfully. An efficient masking technique is proposed to filter all switch allocation requests that are not able to pass flits to the output port, either due to the lack of free space in assigned VC or due to the lack of free VC in the output port for no assigned VC requests. The masking technique also provides an efficient usage of VC memory buffers. The proposed technique has minimal impact in timing and area overhead of a NoC router.

III. CONVENTIONAL NoC ROUTER ARCHITECTURE

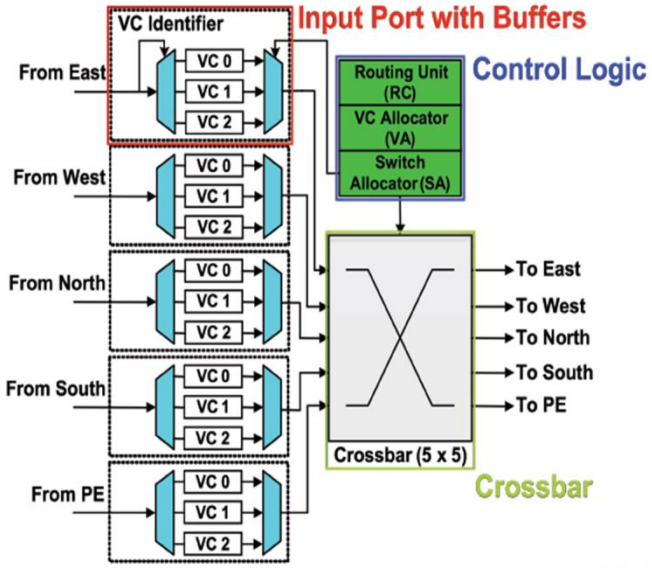
A conventional NoC router has four consecutive pipeline stages.

(1) Route computation: this stage determines the output port that a packet must be sent to.
(2) VC allocation: this stage assigns an empty VC in the neighboring router connected to the output port. Since several header flits may send requests for the same VC, arbitration is required. The routing computation as well as the VC allocation only requires the header flit. The body and tail flits will follow their respective header flit.
(3) Switch allocation: if VC allocation is successful, the third stage sends request to the switch allocator to allocate the output port.
(4) Switch traversal: if the switch allocation is successful, the flit will be passed to the crossbar and be delivered to the output port.

A conventional virtual channel NoC router block diagram is illustrated in Figure 1. The NoC router consists of input ports, VC/SW allocators, routing computation module, and a crossbar. The input ports buffer input flits and send requests to the allocators. The routing computation module determines the output port based on the routing algorithm. After the route computation, a free output VC (OVC) in the next router is assigned to the input VC (IVC) by sending request to the VC allocator. If an OVC is successfully assigned, then another allocation request will be sent to the switch allocator. The crossbar is then configured to send the desired flit to the output port if the switch allocation request is granted. In

order to send requests to the switch allocator, the available space in the next router buffer must be known. Hence, output port modules maintain a set of credit counters to keep track of available buffer space for each OVC.

Figure 1: Conventional NoC Router architecture



The allocator is the most challenging module to design since the overall NoC router performance and area overhead will be dominated by this module. Moreover, allocators are located in the NoC critical path. An allocation is required when several agents (IVCs) require access to several resources (OVCs or output ports) simultaneously. Generally, three types of allocators are widely used in NoC router microarchitecture design, namely, wavefront, separable input-first, and separable output-first allocators.

A comprehensive analysis on the allocators shows that separable input-first allocators have the advantage of lower communication delay, area overhead, and power consumption compared to other schemes. Hence, the separable input-first allocator has been chosen to be implemented in our low latency NoC router. A separable input-first allocator consists of two levels of arbitrations. In the first arbitration stage, for each input port, only one request of all IVC requests is granted. Since several input ports may request the same output resource, another arbitration stage is required to resolve this limitation. The number of arbiters and the arbiters' size required for the VC and switch allocations. VC allocator consumes a large number of resources compared to SW allocator. In this work, we assume that no input port sends a packet.

IV. ANALYSIS OF RESULTS

In order to compare network performance results, we generate cycle-accurate behavioral model of the one-clock-cycle latency CONNECT and our proposed architecture using Verilator. Both NoC are configured in a 5 × 5 mesh topology having 4 VCs on each port with the size of 4 flits per each VC and the flit payload width of 32 bits. As CONNECT [12] only supports dimension order routing (DoR), we use DoR for both routers. All NoC endpoints are connected to the custom traffic generator modules. The traffic generators are responsible for injecting network packets into the NoC routers and collecting performance statistics as the packets are received by destination cores. . The hardware utilization summary and the maximum operating frequency obtained for Altera Cyclone IV EP4CE115 are shown in Table 1.

Table 1: Hardware Utilization summary

4VCs/4 flits per VC	Two clock cycle latency Router	CONNECT Two - Clock	CONNECT One - Clock
Logic cells (LCs)	2,890 (2.5%)	5,934 (5.2%)	5,690 (5.0%)
Memory blocks (M9K)	5(1.2%)	-	-
Maximum frequency	88MHz	65MHz	41MHz

V. CONCLUSION

The paper has given survey of different NoC router designs in terms of hardware utilization such as logic cells, memory blocks and maximum frequency.

VI. REFERENCES

[1]. D. Sanchez, G. Michelogiannakis, and C. Kozyrakis, An analysis of interconnection networks for large scale chip-multiprocessors, ACM Transactions on Architecture and Code Optimization, 7(1):4:1C4:28, 2010.

[2]. Nasim Nasirian ,MagdyBayoumi,"Low-Latency Power-Efficient Adaptive Router Design for Network-on-Chip", IEEE International System-on-Chip Conference (SOCC),2015

[3]. Shalimar Rasheed, Paul V. Gratz, SrinivasShakkottai, Jiang Hu,"STORM: A Simple Traffic-Optimized Router Microarchitecture for Networks-on-Chip", Eighth IEEE/ACM International Symposium on Networks-on-Chip (NoCS),2014

[4]. AlirezaMonemi, Chia Yee Ooi, Maurizio Palesi, Muhammad NadzirMarsono,"Low Latency Network-on-Chip Router Using Static Straight Allocator", International Conference on Information Technology, Computer, and Electrical Engineering ,2016

[5]. P. Gratz, B. Grot, and S. Keckler, "Regional congestion awareness for load balance in networks-on-chip," in High Performance Computer Architecture, 2008, pp. 203–214.

[6]. A. Hansson, K. Goossens, and A.Radulescu, "Avoiding message- dependent deadlock in network-based systems on chip," VLSI design, vol. 2007, 2007.

[7]. Heisswolf, Jan, et al, "Virtual networks–distributed communication resource management," ACM Transactions on Reconfigurable Technology and Systems, vol. 6, no. 2, 2013

[8]. S. T. Nguyen and S. Oyanagi, "The design of on-the-fly virtualchannel allocation for low cost high performance on-chiprouters," in Proceedings of the 1st International Conferenceon Networking and Computing (ICNC '10), pp. 88–94, IEEE,2010.

[9]. Y. Lu, J. McCanny, and S. Sezer, "Exploring virtual-channelarchitecture in FPGA based networks-on-chip," in Proceedingsof the 24th IEEEInternationalSystemonChipConference (SOCC'11), pp. 302–307, 2011.

[10]. M. K. Papamichael and J. C. Hoe, "CONNECT: re-examiningconventional wisdom for designing nocs in the context ofFPGAs," in Proceedings of the 2012 ACM/SIGDA InternationalSymposium on Field Programmable Gate Arrays (FPGA '12), pp.37–46, ACM, 2012.

[11]. Yuho Jin, Ki Hwan Yum,Eun Jung Kim,"Adaptive Data Compression for High-Performance Low-Power On-Chip Networks",International Symposium on Microarchitecture,2008

[12]. R. Akbar , F. Safaei , S. M. SeyyedModallalkar,"A novel power efficient adaptive RED-based flow control mechanism for networks-on-chip",Computer and electrical engineering, Volume 51, April 2016, Pages 121-138

[13]. Cunlu Li, Dezun Dong, Zhonghai Lu, Xiangke Liao,"RoB-Router : A Reorder Buffer Enabled LowLatency Network-on-Chip Router",IEEE Transactions on Parallel and Distributed Systems,2018

[14]. Ranjita Dash , AmartyaMajumdar, VinodPangracious, Ashok Kumar Turuk, Jose L. Risco-Martín,"ATAR: An Adaptive Thermal-Aware Routing Algorithm for 3-D Network-on-Chip Systems",IEEE Transactions on Components, Packaging and Manufacturing Technology , Volume: PP, Issue: 99,2018

[15]. KarthiDuraisamy, YuankunXue, Paul Bogdan, ParthaPratimPande,"Multicast-Aware High-Performance Wireless Network-on-Chip

Architectures",IEEE Transactions on Very Large Scale Integration (VLSI) Systems ,Volume: 25, Issue: 3, 2017

[16]. SergiAbadal , JosepTorrellas, Eduard Alarcon, Albert Cabellos-Aparicio,"OrthoNoC: A Broadcast-Oriented Dual-Plane Wireless Network-on-Chip Architecture", IEEE Transactions on Parallel and Distributed Systems (Volume: 29, Issue: 3, 2018

[17]. A. Karthikeyana, P. SenthilKumar,"Randomly prioritized buffer-less routing architecture for 3D Network on Chip", Computers& Electrical Engineering, Volume 59, Pages 39-50,2017

[18]. Xiao-Wei Shen , Xiao-Chun Ye , Xu Tan , Da Wang ,, Lunkai Zhang , Wen-Ming , Zhi-Min Zhang , Dong-Rui Fan , Ning-Hui Sun," An Efficient Network-on-Chip Router for Dataflow Architecture",Journal of Computer Science and Technology, Volume 32, Issue 1, pp 11–25, 2017

[19]. Nan Su, HuaxiGu,Kun Wang, Xiaoshan Yu, Bowen Zhang," A highly efficient dynamic router for application-oriented network on chip",The Journal of Supercomputing, Volume 74, Issue 7, pp 2905–2915, 2018

[20]. AlirezaMonemi , Jia Wei Tang , Maurizio Palesi, Muhammad N. Marsono,"ProNoC: A Low Latency Network-on-Chip based Many-Core System-on-Chip Prototyping Platform",Microprocessors and Microsystems, Volume 54, Pages 60-74,October 2017

[21]. A. Monemi, C. Ooi and M. Marsono, "Low Latency Network-on-Chip Router Micro architecture Using Request Masking Technique", International Journal of Reconfigurable Computing, vol. 2015, pp. 1-13, 2015