An Efficient Network On-Chip Architecture Based on Isolating Local and non-Local Communications

Vahideh Akhlaghi[1], Mehdi Kamal[1], Ali Afzali-Kusha[1], Massoud Pedram[2]
[1]Nanoelectronic Center of Excellence, College of Engineering, University of Tehran, Tehran, Iran
[2] Department of EE-systems, University of Sothern California, Los Angeles, U.S.A.

***Abstract*—In this paper, we propose a scheme for reducing the latency of packets transmitted via on-chip interconnect network in MultiProcessor Systems on Chips (MPSoCs). In this scheme, the network architecture separates the packets transmitted to near destinations from those transmitted to distant ones by using two network layers. These two layers are realized by dividing the channel width among the cores. The optimum ratio for the channel width division is a function of relative significances of the two types of communications. Simulation results indicate that for non-uniform traffic constituting of more than 30 percent local traffic, the proposed network, on average provides 64% and 70% improvement over the conventional one in terms of average network latency and Energy-Delay product (EDP), respectively. Also, for uniform and NED traffic patterns, by adjusting the number of hops between local nodes to include approximately 55 percent of total communications in local ones, the proposed architecture provides the latency reduction of 50%.**

I. INTRODUCTION

To improve performance and power consumption of applications running on an electronic device, modern embedded systems need to be implemented on a single chip. Also, continuing advances in integrated circuit fabrication technologies has enabled the use of multiprocessor system-on-chip (MPSoC) architectures [1]. In the state of the art integrated circuits, the delay and energy consumption of (semi-) global interconnects are considerably larger than those of elementary operations performed in logic gates. Therefore, the performance of MPSoCs strongly depends on the underlying on-chip interconnect. To improve the communication performance, the traditional shared bus architecture may be replaced by Network-on-Chip (NoC) providing simultaneous transmission of packets. To ensure the quality of service, virtual channels (VCs) switches have been proposed [2]. Owing to logical resources needed to perform virtual channels allocation in hardware, this technique presents serious complexity problems [3]. This issue can be reduced by using multiple physical networks instead of VCs [4][5].

In addition, to reduce the communication latency and energy consumption, task mapping algorithms has increased communication locality by placing processing cores which considerably communicated with each other as close as possible to each other on the chip [6][7]. Several locality-aware on-chip communication schemes have been proposed in the literature [8][9]. A locality-aware network topology has been presented in [8], which addresses the communication locality issue by introducing two levels of networks: local and global. The local network is a shared bus while the global network is a low radix mesh connecting local networks. Asymmetric buffer assignment is another locality-aware NoC architecture where a larger buffer is assigned to the core port while smaller buffers are considered for other ports of the router [9]. Also, two inter-processor interconnect schemes using multiple links are compared in this work. One scheme is based on separating the nearest-neighbor links and long distance links, whereas the other utilizes each link for both local and non-local communications.

In designing a NoC, link optimization is of critical importance. Because wide links lead to increase in the routing resources needed (viz. the crossbar area increases quadratically with port width). Also, routing difficulties due to the limited number of metal layers impose limitation on the number of wires that can be successfully routed [10]. In [10] two *b*-bit unidirectional links between two routers are replaced by one *b*-bit bidirectional link being split into *n* channels with the same width. Besides, one flit is divided into smaller units, called phit (physical transfer unit). So, multiple phits can be transmitted via multiple channels of width *b*/*n*.

In this work, we propose an efficient locality-aware on-chip interconnect architecture where separate local and non-local communications are performed by dividing the channel width between two network layers. The proposed architecture exhibits higher efficiency of multiple physical networks over virtual channel switches. Also, our design takes advantage of dividing a link's width over using multiple links, thereby, keeping the number of wires the same as that of the baseline NoC. Finally, we define the locality based on the distance between the source and destination nodes so that for any traffic profile, the proposed architecture leads to the lower latency.

The rest of this paper is organized as follow. The motivation behind this work is brought in Section II. The detailed design of the proposed router architecture is described in Section III. The simulation results are discussed in Section IV. Finally, Section V concludes this paper.

II. MOTIVATION

In this section, the importance of separating the local and non-local data communication in a NoC is investigated using synthetic traffic profiles. Let us consider the synthetic traffic profile of Negative Exponential Distribution (NED) [11] in which the communication probability decreases exponentially as distance between two nodes increases. For this profile, if local communications are defined based on a distance of one hop or at most two hops, on average, for the 5×5 mesh, they form 30% and 60% of the total communications, respectively.

978-3-9815370-0-0/DATE13/©2013 EDAA

Since these numbers are large enough, one may expect that some latency improvement may be achieved by separating local and non-local communications. As another synthetic traffic profile, consider uniform traffic where each node sends data to other nodes with an equal probability. Now, in the case of 5×5 mesh, on average, single hop local communications are about 12 percent of the total communications which does not justify the use of a separate layer. By increasing the distance to either at most two hops or three hops, local communications become 34 percent and 57 percent of total communications, respectively, which makes it justifiable to use separate network resources for local and non-local communications.

III. The Communication Architecture

A conventional router includes five input and five output ports (North, West, South, East, and Local). In NoCs, the source node injects packets into the network. The routing computation (RC) unit determines the path of an incoming packet based on the destination address in its header flit. Then, the flits of the packet are directed to the appropriate output Physical Channel (PC) to go to the next switch. Since more than one packet may simultaneously request for the same physical channel, there are a multiplexer (MUX) and an arbiter at each output port to select one packet at each time. Our proposed network architecture, shown in Fig. 1, is based on the conventional one. In the proposed architecture, the physical channel width is divided to create two separate network layers. One layer is used for local communication while the other is used for non-local one. We denote the two layers of networks as layer *A* (for local communications) and layer *B* (for non-local communications).

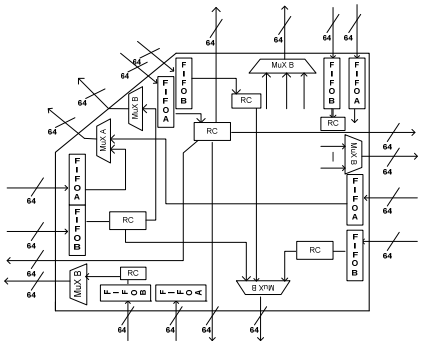

Fig. 1. The proposed communication architecture (local communication is single hop). The physical channel width (128 bit here) is divided (both widths are 64 bits here).

The architecture of layer *B* is the same as that of the conventional network. The hardware complexity of layer *A* depends on the definition of local communications. If the local communication is defined as that established between two immediate neighboring nodes, logical complexities of layer *A* are lower than those of layer *B*. The reason is that Arbiter and RC units are no longer needed for other ports of layer *A* except for the local one. This is because only the local port may request the North, West, South and East ports to use their physical channels. So, the Mux and Arbiter unit of four cardinal output ports can be eliminated. Also, RC unit is not needed to calculate the route of local packets coming from East, North, West, or South ports since all of them are destined for the core connected to the switch. If the local communication is defined as more than one hop connection, logical complexities of layer *A* will be the same as those of layer *B*.

We denote the inter-processor communication architecture by BDNoC (Bitwidth Division in Network on Chip). This network is fully specified using three parameters of *x*, *y*, and *z* (BDNoC [(*x*,*y*),*z*]). The parameters *x* and *y* denote the widths devoted to layers *A* and *B*, respectively. The parameter *z* indicates the maximum number of hops allowed for local communications.

IV. Results and Discussion

A. Simulation Setup

In order to assess the efficiency of the proposed network architecture, we implemented the actual hardware of both conventional and BDNoC architectures using VHDL, and obtained *average network latency* (ANL) by calculating the number of clock cycles needed for a packet to reach from a source to a destination. The reported ANL in this section is the true ANL divided by 100. In simulations, each node sent 5,000 packets to other nodes under synthetic traffics. Each packet had 512 bits which was broken into a number of flits. The widths of channels between routers were 128 bits, and the number of flits was 4. When a link was divided into two links, the number of flits changed. In this study, the number of flits was varied from 5 to 22 (assuming that the width of narrower links ranges from 104 to 24). Routers used the deterministic *X-Y* routing algorithm, input buffering, and wormhole switching.

B. The Division Ratio

In this section, for several percentages of local communications, several width division schemes were tested to determine the optimum division ratio for 5×5 BDNoC with 128-bit links. We assume that the traffic pattern is non-uniform/localized (*x*) [8]. Let us denote the local traffic percentage that each node of the mesh sends (uniformly) to its immediate neighboring nodes by *x*. The rest of packets (i.e. (100-*x*) percent) are uniformly sent to other nodes. The BDNoC architecture provided performance improvements compared to a conventional NoC for different values of *x* ranging from 30 to 80. We do not study locality rates of above 80 since it is not reasonable to use a separate layer for more than 80 percent traffic. Also, using BDNoC for above 70 percent non-local communications due to traversing more hops through narrower links worsen the latency of a packet.

Fig. 2 shows the ANL of BDNoC and conventional NoC under different Packet Injection Rates (PIR). Due to limited space, we just bring the plots for Non-uniform (30), (50) and (70). The set of division ratios which improves the latency of the proposed network over conventional one is marked on each figure by an oval. As is expected, when the rate of the locality increases, larger widths should be devoted for layer *A* (the oval moves to the right) to achieve improvement.

Fig. 3 depicts the efficacy of the proposed NoC under different PIRs. It is observed that the conventional architecture

with over 30% local traffic saturates at a smaller PIR than the proposed one. Also, Fig. 3(b) shows that although the ratios of local and non-local communications are the same, the division ratio of (64-64) is not as efficient as (40-88) and (48-80). The reason is that non-local packets traverse more hops than local ones. Hence, one should devote a larger fraction of the channel width to the non-local communications. In addition, at higher PIRs, (40,88) leads to better performance improvement than (48,80). It originates from the fact that, at higher injection rates, the probability of traffic congestion for non-local packets which passes through more hops increases more compared to the local ones; so, the non-local width part should be increased.

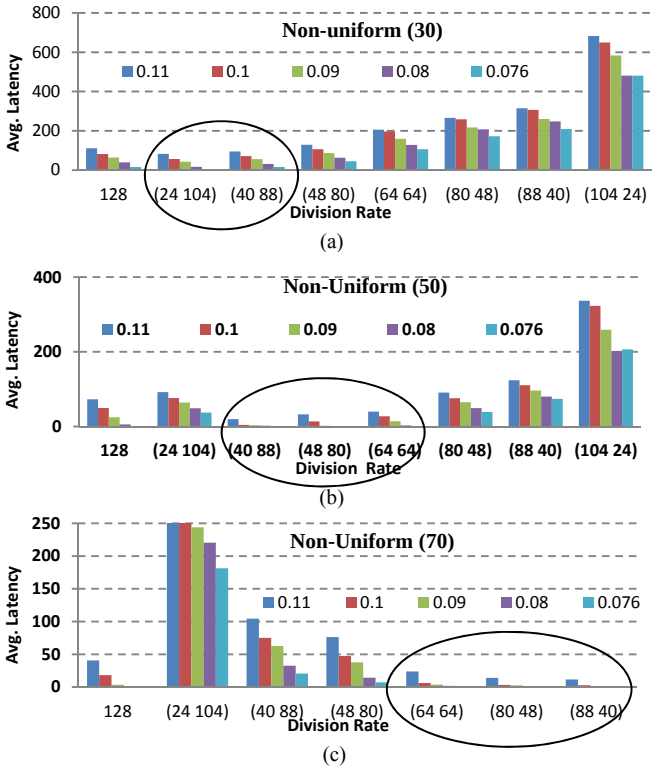

Fig. 2. Average network latency of BDNoC [(*x*,*y*),1] and conventional NoC with varying packet injection rates.

C. Optimum Local Distance for BDNOC

Here, we show how to define local communications based on the number of hops so that BDNoC results in the smallest values in ANL under different traffic profiles. For this purpose, we compare the efficacy of the proposed network architecture with different definitions of local communications under uniform and NED traffic distributions.

Fig. 4 provides ANL of a 3×3 mesh under NED traffic for channel widths of 128 bits. It shows that BDNoC [(64,64),1], BDNoC [(40,88),1] and BDNoC [(104,24),2] give rise to lower latency than the conventional NoC. Also, results suggest that defining local packets as those which traverse one hop provides better improvement since local communications are 50 percent for this case (*i.e.*, balanced workloads on two layers). Fig.5 provides the ANL of a 5×5 mesh under NED traffic distribution for channel widths of 128 bits. Again, the figure reveals that BDNoC [(80,48),2] which comprises 60 percent of the whole communications as local ones results in considerable performance gain. Next, ANL of the 5×5 mesh under Uniform traffic is obtained and shown in Fig. 6. It indicates that BDNoC [(*x*,*y*),3] leads to higher performance, mainly because local communications are 57%.

Based on the above discussion, including around 55 percent of NED and Uniform traffics in the local traffic, BDNoC can lead to significant improvement (*i.e.*, 50%, on average) in terms of ANL.

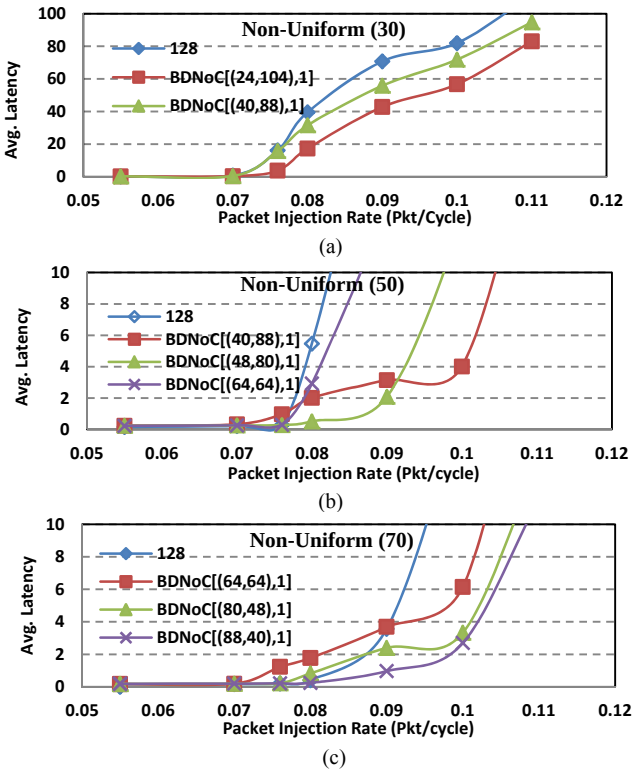

Fig. 3. Average latency versus packet injection rate of 5×5 mesh with different non-uniform traffics for one-hop local communication.

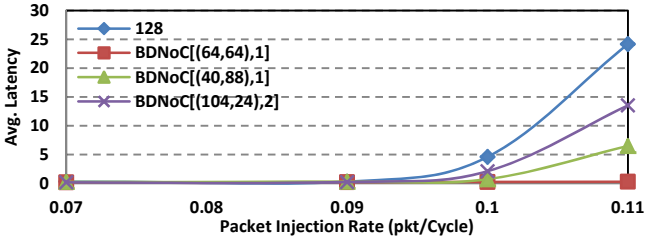

Fig. 4. Average network latency of a 3×3 mesh with NED traffic.

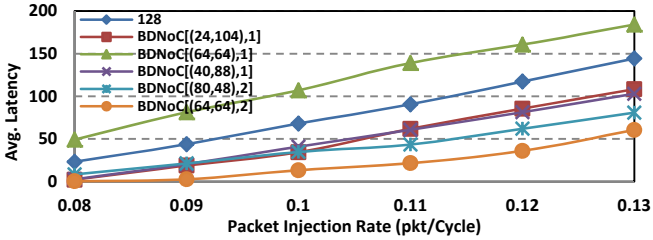

Fig. 5. Average network latency of a 5×5 mesh with NED traffic.

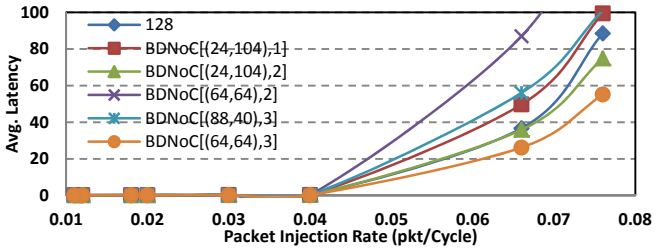

Fig. 6. Average network latency of a 5×5 mesh with Uniform traffic.

D. Energy-Delay Product

In this part, we assess the efficiency of the proposed router architecture in terms of Energy-Delay product (EDP). For this purpose, the VHDL models of the BDNoC and conventional NoC architectures were synthesized using a 45nm standard CMOS Library [12]. The simulation results were obtained from Modelsim and the generated VCD files were fed into the Synopsys PrimePower tool to compute the power consumption. The average power dissipation of the two routers and each layer of BDNoC are presented in Table I. Note that the power of wires has not been considered in the power calculation. By multiplying ANL of the conventional architecture by its power, energy consumption of a packet can be calculated. Also, energy consumption of BDNoC was obtained using the following expression:

$$Energy=\alpha.P_A.Lat_{Avg,A}+(1-\alpha).P_B.Lat_{Avg,B} \quad (1)$$

Here, α is the percentage of the local communications, P_A and P_B are the power consumptions of layers A and B, and $Lat_{\mathrm{Avg},A}$ and $Lat_{\mathrm{Avg},B}$ are the average network latencies of the layers A and B, respectively.

TABLE I. Power Consumption of Each Router.

Router	Total Power Consumption (mW)
BDNoC [(*x*,*y*),1]	41.7
ConvNoC	19
Layer *A* of BDNoC [(*x*,*y*),1]	20.97
Layer *B* of BDNoC [(*x*,*y*),1]	20.77

Fig. 7 provides EDP plots of BDNoC [(*x*,*y*),1] and the conventional NoC for different rates of local communications versus PIR. The total link width was 128. In calculating the EDP in BDNoC, for each locality rate, we chose the width division ratio leading to the better performance improvement. In general, at PIRs above 0.08, the proposed architecture outperforms the conventional one in terms of EDP. Particularly, when the local communication rate is 50%, a significant improvement is achieved owning to the balanced traffic distribution over the two network layers *A* and *B*.

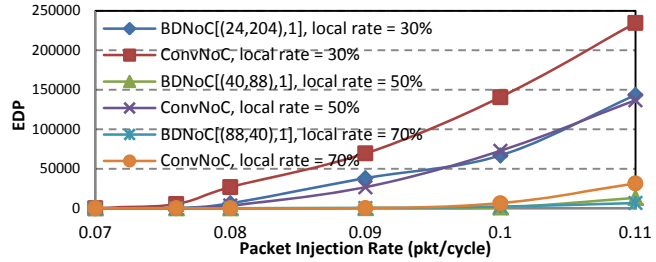

Fig. 7. Energy delay product for different non-uniform traffics.

V. Conclusion

A locality-aware on-chip interconnect architecture where local and non-local communications were carried out separately by dividing the physical channel width was proposed. In this work, the local communications were defined based on the number of hops that they passed through. We compared the efficacies of the proposed and the conventional router architectures for different percentages of local communications. The results showed that if the link width was 128, for the local communication rates above 30 percent, there were division ratios that give rise to on average 64% reduction in the network latency. Additionally, we evaluated the performances of the proposed NoC under different definitions of local communications under the NED and uniform traffic patterns. The results indicated that defining local communications such that they formed approximately 55 percent of all the communications would lead to maximum reduction (50%, on average) in the packet latency. Furthermore, we studied the Energy-Delay product of a packet transmitted through the proposed routers and conventional ones. It was found that the proposed network improves EDP by 70% compared to the conventional one.

References

[1] M.B. Taylor, *et al.*, "A 16-issue Multiple-program-counter Microprocessor with Point-to-point Scalar Operand Network", in Proceedings of International Solid-State Circuits Conference, 2003, pp. 170–171.

[2] W.J. Dally, "Virtual-channel Flow Control," in Proceedings of ACM/IEEE International Symposium on Computer Architecture (ISCA), 1990, pp. 60-68.

[3] N.Banerjee *et al.*, "A Power and Performance Model for Network-on-Chip Architectures," in Proceedings of ACM/IEEE DATE, 2004, vo. 2, pp. 1250-1255.

[4] F. Gilabert, *et al.*, "Improved Utilization of NoC Channel Bandwidth by Switch Replication for Cost-effective Multi-processor Systems-on-Chip," in Proceedings of ACM/IEEE NOCS, 2010, pp. 165–172.

[5] M.R. Kakoee, *et al.*, "ReliNoC: A Reliable Network for Priority-based On-chip Communication," in Proceedings of ACM/IEEE DATE, 2011, pp. 1-6.

[6] E. Carvalho and F. Moraes, "Congestion-aware Task Mapping in Heterogeneous MPSoCs," in Proceedings of International Symposium on System-on-Chip (SoC), 2008, pp. 1-4.

[7] A.K. Singh, *et al.*, "Efficient Heuristics for Minimizing Communication Overhead in NoC-based Heterogeneous MPSoC Platforms," in Proceedings of International Symposium on Rapid System Prototyping, 2009, pp. 55-60.

[8] R. Das, *et al.*, "Design and Evaluation of a Hierarchical On-Chip Interconnect for Next-Generation CMPs," in Proceedings of IEEE HPCA, 2009, pp. 175-186.

[9] Zhiyi Yu, B.M. Baas, "A Low-Area Multi-Link Interconnect Architecture for GALS Chip Multiprocessors," IEEE Transactions on Very Large Scale Integration Systems, vol. 18, pp. 750-762, 2010.

[10] R. Hesse, *et al.*, "Fine-Grained Bandwidth Adaptivity in Networks-on-Chip Using Bidirectional Channels", in Proceedings of ACM/IEEE NoCS, 2012. pp. 132-141.

[11] A-M. rahmani, *et al.*, "NED: A Novel Synthetic Traffic Pattern for Power/Performance Analysis of Network-on-chips Using Negative Exponential Distribution," J. of Low Power Electronics, vol. 5, pp. 396-405, 2009.

[12] FreePDK, A Free OpenAccess 45nm PDK and Cell Library for university, http:// www.eda.ncsu.edu.