ARTICLE IN PRESS

Computers and Electrical Engineering xxx (2014) xxx–xxx

Contents lists available at ScienceDirect

Computers and Electrical Engineering

journal homepage: www.elsevier.com/locate/compeleceng

An efficient network on-chip architecture based on isolating local and non-local communications ☆

Vahideh Akhlaghi [a], Mehdi Kamal [a], Ali Afzali-Kusha [a,*], Massoud Pedram [b]

[a] Nanoelectronic Center of Excellence, College of Engineering, University of Tehran, Tehran, Iran
[b] Department of EE-Systems, University of Sothern California, Los Angeles, USA

ARTICLE INFO

Article history:
Received 27 February 2014
Received in revised form 1 December 2014
Accepted 3 December 2014
Available online xxxx

Keywords:
Network-on-Chips
Interconnect architecture
Communication locality
MultiProcessor Systems on Chips

ABSTRACT

In this paper, a locality aware NoC communication architecture is proposed. The architecture may reduce the energy consumption and latency in MultiProcessor Systems on Chips (MPSoCs). It consists of two network layers which one layer is dedicated to the packets transmitted to near destinations and the other layer is used for the packets transmitted to far destinations. The actual physical channel width connecting the cores is divided between the two layers. The locality is defined based on the number of hops between the nodes. The relative significances of the two types of communications determine the optimum ratio for the channel width division. To assess the efficiency of the proposed method, we compare its communication latency with that of conventional one for different channel widths, communication traffic profiles, and mesh sizes.

1. Introduction

Technological advances in the integrated circuit fabrication have provided an opportunity to implement embedded systems on a single chip. It has led to the advent of MPSoC architectures which are advantageous to single processors in terms of performance and energy consumption. These architectures have been used in a wide range of applications including home and mobile multimedia, biomedical, automotive and networking [1,2].

In the state of the art integrated circuits, logic gates perform the operations faster, whereas RC delay of (semi-) global interconnects becomes longer. Therefore, in realizing high performance and energy-efficient Chip Multiprocessor (CMP) systems, an efficient underlying on-chip interconnect architecture is crucial. To improve data communication parameters such as delay and energy consumption, Network-on-Chip (NoC) architecture has emerged as a substitution for the traditional shared buses. NoC benefits from simultaneous transmission of packets between the cores. This architecture which routes packets instead of wires [3,2,4], provides higher degree of scalability and reusability compared to the shared bus.

For these chips, tasks are mapped onto cores. Modern task mapping algorithms (viz. Nearest Neighbor (NN) and Best Neighbor (BN)) increase communication locality in today MPSoCs to reduce the communication costs. In NN [5], the first task is mapped on the first Intellectual Property (IP) block as an initiator. Then, the next task from the task graph is mapped to the nearest available block. BN determines the next block according to the nearest neighbor paradigm and Path Load (PL) algorithm [5,6]. PL is a congestion-aware algorithm to reduce the occupancy of channels used by the tasks which are mapped.

☆ Reviews processed and recommended for publication to the Editor-in-Chief by Associate Editor Dr. Masoud Daneshtalab.

* Corresponding author.
E-mail addresses: vahide.akhlaghi@ut.ac.ir (V. Akhlaghi), mehdikamal@ut.ac.ir (M. Kamal), afzali@ut.ac.ir (A. Afzali-Kusha), pedram@usc.edu (M. Pedram).

http://dx.doi.org/10.1016/j.compeleceng.2014.12.002
0045-7906/

The technique introduced in [7] divides the entire NoC into virtual clusters. The blocks of each cluster are dedicated to the tasks of one application which again results in the locality of communication.

In this paper, to improve the parameters of the communication architecture of NoCs, distinguishing between local and non-local communications, we make the following contributions:

- First, we introduce low-latency yet energy-efficient two-layer on-chip interconnect architecture which separate local and non-local communications using two different layers. In order to avoid adding a separate physical channel between the cores, the channel width is divided between the two network layers.
- Second, in this work, the effect of varying the number of hops is studied to define the local communication on the communication parameters for different traffic profiles.

The rest of this paper is organized as follows. Section 2 describes the motivation behind proposing locality-aware network on chip and Section 3 reviews the related works on the areas of locality-aware communication and link optimization of NoCs. The structures of the generic and proposed router architectures are explained in Section 4. The simulation setup and results are discussed in Section 5. Finally, Section 6 concludes the paper.

2. Motivation

In this section, we use synthetic traffic profiles to give the motivation for separating the local and non-local data communications in a NoC. A novel synthetic traffic profile which considers increased local communication is Negative Exponential Distribution (NED) [8]. In NED, the probability of data communication between the nodes depends exponentially on their physical distance (*i.e.*, number of hops) in the interconnect network where the probability decreases by increasing the distance. This synthetic traffic profile may resemble a more realistic situation if the mapping focuses on increasing local communication. For the case of a NoC with a mesh size of 3×3 under the NED traffic profile, the probability of sending packets from a source node to a destination node that is n hop farther is shown in Table 1. In the leftmost column, a pair of numbers represents the position of a node in the mesh. The first number indicates the row number whereas the second one shows the column number. The probabilities are obtained by solving the equations given in [8]. For this profile, if the local communications are defined based on a distance of one hop, on average, for the mesh sizes of 3×3 and 5×5, they form 50% and 30% of the total communications, respectively. These numbers indicate that local communication form a considerable fraction of communication volume inspiring us for separating local and non-local communications across network on chips.

As another synthetic traffic profile, we investigate communication locality in the Uniform traffic pattern where the packets are sent to each node with an equal probability. Our results show that even for this type of traffic profile, an acceptable percentage of the total communications may be considered as local ones justifying a separate resource allocation. For this traffic profile, in the case of 5×5 mesh, defining local communications based on a single hop separation between the nodes constitutes about 12% of the total communications which is not large enough for using a separate communication layer. In the Uniform traffic pattern, each node sends packet to other nodes uniformly. In the case of 5×5 mesh with the locality definition of one hop, there are 25 nodes sending 1/24 of its packets to another node in a mesh. Since the nodes located in the borders have 2 or 3 neighboring nodes while other nodes have four neighbors, the average number of local nodes turns out to be about 3. Therefore, on average, about 1/8 of packets become local ones (*i.e.*, 12% local traffics). By increasing the number of hops for local communications to up to two hops, this type of communication forms 34% of the total communication. If communication locality is defined based on up to three hops between the source and destination nodes, they will comprise 57% of the whole communications. In the latter two cases, the use of separate network resources for local and non-local communications is justifiable.

Therefore, based on the percentage of local communication mentioned earlier, it makes sense to suggest a NoC architecture which employs two network layers, one for local and one for non-local communications. As is discussed in Section 5, the results of our investigation show that this architecture can lead to decrease in the packet latency.

3. Related works

There are many research works in the literature devoted to NoC architecture and routing algorithms. The focus of this section is on reviewing works on multilayer networks on-chip, locality-aware on-chip communication, and link utilization techniques.

Table 1
Probability of sending packets in a 3×3 mesh under NED Profile.

Source Node	Dist. of 1 hop	Dist. of 2 hops	Dist. of 3 hops	Dist. of 4 hops
(0,0)	0.42	0.38	0.15	0.05
(0,1)	0.56	0.32	0.12	–
(1,1)	0.65	0.35	–	–

Please cite this article in press as: Akhlaghi V et al. An efficient network on-chip architecture based on isolating local and non-local communications. Comput Electr Eng (2014), http://dx.doi.org/10.1016/j.compeleceng.2014.12.002

For the purpose of maintaining the Quality of Service (QoS), Virtual Channels (VCs) switches have been proposed (see, *e.g.*, [9–12]). In this scheme, QoS traffic classes use high priority VCs whereas normal classes utilize low priority VCs. Since the implementation of the technique requires a virtual channel allocation hardware unit, the complexity of the design will be higher [11]. Using multiple physical networks instead of VCs can reduce the complexity while performance improvement and supporting separate traffic classes may be achieved (see, e.g., [13–15]).

Several locality-aware hardware and software based techniques for improving the latency and energy consumption of on-chip communication have been proposed in the literature (see, *e.g.*, [16–18]). In [16], the authors proposed a locality-aware network topology to minimize the energy delay product of NoCs. The topology consisted of two levels of networks: the local ones which were shared buses connected together via a global level which is a low radix mesh. Local communications use the local level while non-local communications are performed through the mesh. Another approach exploiting locality in on-chip communication has been suggested in [17]. In this scheme, each node in the array of processors communicates only with its four immediate neighboring nodes. This platform is used for DSP application, and avoids long distance communication by complicating intermediate processors. The technique presented in [18], considered the idea of locality aware NoC architecture by presenting asymmetric buffer assignment. In this method, a larger buffer was assigned to the core port which was assumed to be the destination of most of incoming packets, while the smaller buffers were considered for the other ports of the router. Also, the efficiencies of two different schemes using multiple links between two switches were compared in this work. One scheme separated nearest-neighbor and long distance links, whereas the other employed each link for both local and non-local communications. Both schemes are effective while the cost of the first one was lower while the second one was more flexible.

In addition, the full utilization of links is of critical importance in designing a NoC. The reason is that using wider links calls for more routing resources (*e.g.*, the crossbar area is quadratically proportional to the port width). Also, the number of wires on a chip is limited due to the fixed number of metal layers [21]. The problem is expected to be intensified with the growing size of SoCs (and increasing the number of processors on a chip). Authors in [19] designed bidirectional interconnects to allow simultaneous transmitting and receiving data using a single interconnect segment of a long bus. Sharded router architecture described in [20] was based on bandwidth partitioning and stealing techniques to provide a full bandwidth utilization. It uses four subnetworks by dividing 128-bit links into four 32-bit links. Since in conventional VC switches, there are four VCs, each of these subnetwroks is dedicated to one VC. In [21], two unidirectional links were replaced by one bidirectional link which was split up into n channels with the same width. To decouple flit width from the channel width, one flit was divided into smaller units, called phit (physical transfer unit). Subsequently, instead of sending one flit over a b-bit channel, multiple phits can be transmitted through n channels of width b/n. An Adaptive Physical Channel Regulator (APCR) for NoC routers was proposed in [22]. By reducing the size of a flit to be less than that of a phit, an APCR router allows flits from different packets to share the same output channel in a single cycle. In [23], SDM (Spatial Division Multiplexing) router for NoC was proposed in which the width of a link was divided. The proposed SDM router used the recursive Benes switch consisting of 2×2 atomic switches (providing direct and cross connections) and two intermediate $(n/2) \times (n/2)$ Benes switches for interconnecting any groups of wires at the input port to other groups at the output port. To decrease the bandwidth demands of a system, a routing algorithm for routing table based NoC routers was introduced in [24]. In order to fully utilize the channels of NoC, Time-Division-Multiplexer routing technique was used. In addition, the banker algorithm was employed to allocate communication resources.

In this work, we propose an efficient interconnect architecture which separates local and non-local communications for an efficient use of available physical link width (bandwidth) without considerably increasing the hardware complexity. The design, which keeps the number of wires the same as the baseline NoC, makes use of multiple physical network technique instead of using virtual channel approach. The details of the proposed architecture will be described in the next section.

4. The communication architecture

4.1. The conventional architecture

A 2D $M \times N$ network on chip with the mesh topology consists of $M \times N$ tiles located in M rows and N columns. Fig. 1 shows a conventional inter-processor communication architecture which has five input and five output ports (North, West, South, East, and Local). In this figure, the details for the East port are provided. Each Processing Element (PE) (*i.e.,* IP blocks or memory unit) is connected to the Local ports of the corresponding tile of NoC through a network interface. The width of the physical channel between NoC tiles may have different values including 32, 64, and 128.

In NoCs, the source node packetizes the data to be transmitted to the destination node and injects packets into the network. The Routing Computation (RC) unit determines the path that the header flit of an incoming packet should take to reach to the destination node. Once the path is determined through the header flit, the other flits of the packet follow the same path by passing through appropriate switches through appropriate output ports connected to Physical Channels (PCs). To compensate for the delay induced by wires between switches, routing computation parts, and link congestions, buffers are inserted at each input port or output port [25]. There may be more than one packet simultaneously requesting a physical output channel. Thus, for each output port, there are a Multiplexer (MUX) and an arbiter unit that determine which packet can use the channel.

Please cite this article in press as: Akhlaghi V et al. An efficient network on-chip architecture based on isolating local and non-local communications. Comput Electr Eng (2014), http://dx.doi.org/10.1016/j.compeleceng.2014.12.002

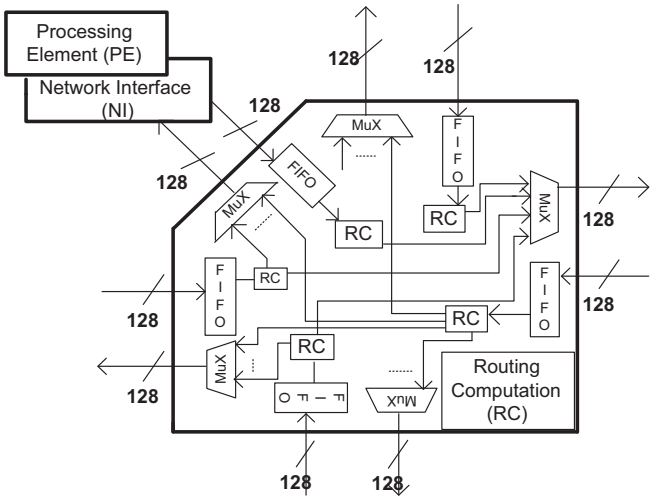

Fig. 1. Conventional communication architecture with a channel width of 128 bits.

4.2. The proposed architecture

In this work, we propose a NoC architecture based on the conventional architecture which is shown in Fig. 2. In this architecture, the physical channel width is divided into two physical network layers which one layer is used for local communication and the other is utilized for non-local communication. The two layers of networks are denoted by layer *A* (local layer) and layer *B* (non-local layer). The logical elements of the layers *A* and *B* are tagged with *A* and *B*, respectively. The architecture of the layer *B* is the same as that of the conventional network. The hardware complexity of the layer *A* depends on the definition of local communications based on the distance between nodes.

If the local communication is considered as the packets which traverse between the nodes with one hop distance, the area and logical complexity of the layer *A* are lower than those of the layer *B*. The reason is that, as shown in Fig. 2, in the layer *A*, other ports except the Local one no longer need the RC and Arbiter units. Because only the Local port may request the North, West, South and East (cardinal) ports to use the corresponding physical channel. The other ports only need to pass the

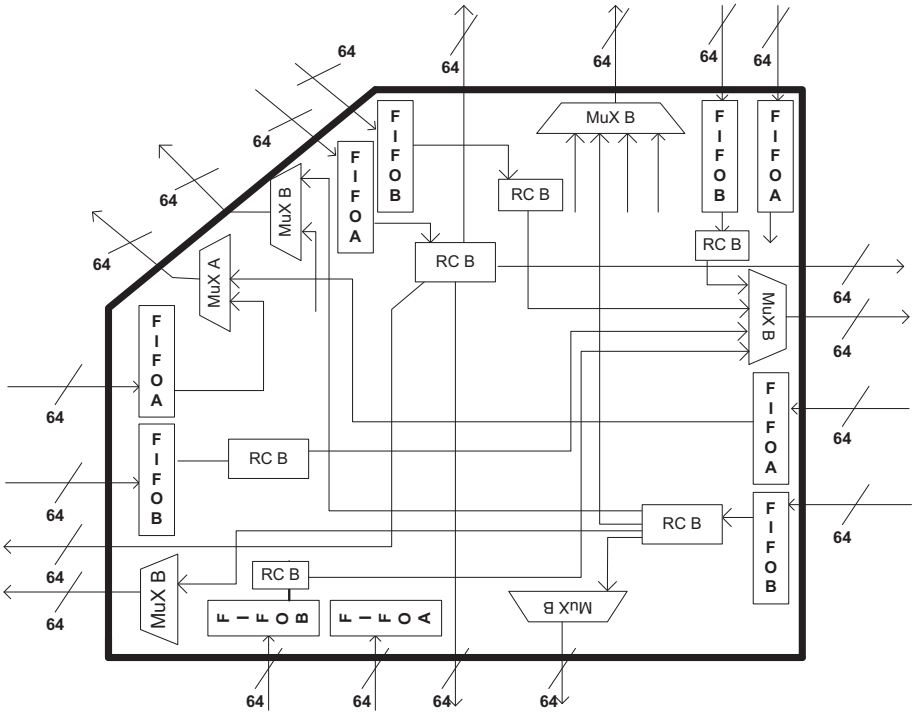

Fig. 2. The proposed communication architecture (assuming local communications are between immediate neighboring nodes). The physical channel width (128 bits here) is divided between two network layers with different widths (assuming two widths of 64 bits in this case).

Please cite this article in press as: Akhlaghi V et al. An efficient network on-chip architecture based on isolating local and non-local communications. Comput Electr Eng (2014), http://dx.doi.org/10.1016/j.compeleceng.2014.12.002

injected packet (coming from the processing core connected to the switch) to the next switch. For instance, since the layer *A* is only employed for transmitting data coming from the processor connected to, *e.g.*, the south neighboring node to the local node, there will not be any request from cardinal input ports for sending data through any of the other cardinal output ports. However, any of the cardinal output ports may be used by the Local port to send the data to any of the immediate neighbors. Hence, there is only one request to each four cardinal output ports, and consequently their MUX and Arbiter can be eliminated. In addition, for the local communications, the packets which are entered into the switch through the East, North, West, or South ports should go to the core connected to the switch, and hence, do not need any routing. Note that, only for the East port, more details of the architecture are drawn in Fig. 2.

If the local communication is defined for the data transfer between the nodes with more than one hop (*e.g.*, two or three hops) distance, the area and logical complexity of the layer *A* are the same as those of the layer *B*.

In the proposed architecture, there are ten input ports and ten output ports. This is realized by spatially dividing the physical channel of the conventional architecture into two channels with lower widths (*e.g.*, 128 bits is split into 48 and 80 bits). Therefore, the width of the physical channel of each port is less than that of the corresponding channel in the conventional architecture. The optimum division ratio is a function of the ratio of the local and non-local communication rates. This architecture should be configured statically at the design time fixing the width of local and non-local sub-links for each percentage of the local communication. By choosing a proper division ratio, the architecture can still outperform the baseline NoC for a larger range of local traffics. This will be discussed in more detail in section V.B.

In this work, the proposed communication architecture is denoted by BDNoC (Bitwidth Division in Network on Chip), which can be fully specified using three numbers of *x*, *y*, and *z* (BDNoC [(*x*,*y*),*z*]). The parameters *x* and *y* show the widths of the channels in the layer *A* and layer *B*, respectively. The maximum number of hops between source and destination nodes in local communications is indicated by the parameter *z*.

Finally, as mentioned before, the purpose of suggesting this technique is to lower the average network latency of the packet transmission. The technique focuses on the underlying communication network. If the latency improves, the operating frequencies of the processing elements, which transmit/receive data, are not deteriorated. In the case of routers, dividing the channel width, decreases their complexities, and hence, their operating frequencies are not lowered. Also, when the latency decreases, the bandwidth utilization should have been increased. Also, a lower latency implies a higher throughput. Based on these explanations, in the cases that the BDNoC technique outperforms the conventional NoC, better bandwidth utilizations and higher throughputs should have been achieved for the proposed approach.

5. Results and discussion

5.1. Simulation setup

To assess the efficiency of the proposed network architecture, the hardware realizations of both conventional and BDNoC architectures were implemented using VHDL. The implementations were used to calculate the average network latency (*Avg. Latency*) of the packet in terms of the number of clock cycles that a packet needs to reach from a source to a destination. The reported average network latency in this section has been normalized to 100 (*i.e.*, divided by 100). In this work, to connect routers, the NoC mesh topologies with the sizes of 3×3 and 5×5 were used. Each node was simulated in the mesh sending 5000 packets to the other nodes under synthetic traffic patterns. The deterministic $X - Y$ routing algorithm and input buffering switches were employed in this work.

Three synthetic traffic patterns, namely Uniform Random (UR), Non-Uniform/localized traffic (NU) [16] and NED were employed to evaluate the performance of the proposed network architecture. In our simulations, each packet had 512 bits and the widths of the channels between routers were 64 and 128 bits. Each packet consisted of several flits. In the case of 128-bit (64-bit) channels, the number of flits was 4 (8). When a link was divided into two links, the number of flits is changed based on the width of the divided link (for example, if 128-bit link was divided into two 40 and 88-bit links, the number of flits would be 13 and 6, respectively). For the division ratios considered in this study for the link width of 128 bits (64 bits), the number of flits varied from 5 (11) to 22 (32) (assuming that the width of divided links ranged from 104 (48) to 24 (16)). The study was performed for different network sizes, channel widths, definitions of local communication and traffic patterns. Finally, the performance was measured using average network latency parameter.

5.2. Determination of division ratio based on locality rate and local distance under non-uniform traffic profile

In this section, the optimum division ratio of a 5×5 BDNoC for several percentages of local communications by testing different width division schemes is determined. The width of the channels among routers was considered to be 64 bits and 128 bits. Here, the traffic pattern which was considered here was non-uniform/localized specified by the parameter *t*. The parameter *t* indicates the percentage of local traffic which is sent uniformly to the neighboring nodes of each node in the mesh. The rest of the packets (*i.e.*, $(100 - t)$% of the traffic) are sent to the non-local nodes of the transmitter node uniformly. For the local communications, it is assumed that one or (at most) two hops between the source and the destination nodes. In the study, several non-uniform traffic distributions including non-uniform (30), (40), (50), (60), (70) and (80) for the 5×5 mesh NoC were considered. Using BDNoC for traversing above 70% non-local communications (passing through more

Please cite this article in press as: Akhlaghi V et al. An efficient network on-chip architecture based on isolating local and non-local communications. Comput Electr Eng (2014), http://dx.doi.org/10.1016/j.compeleceng.2014.12.002

hops than local communication) worsen the latency of a packet compared to that of the conventional one. The locality rates of above 80% were not studied here since it did not make sense to divide the physical channel (considering its area overhead).

The results of the study are shown in Figs. 3–8. The percentage of the local communications for each set of results is shown on top of the figures.

64-bit links -First, the efficiency of the proposed router architecture when the channel width is 64 bits is considered. Fig. 3 shows the comparison of the average latency between BDNoC and conventional NoC when local communication is defined based on one hop. Similar results for the case when at most two hops were used to define the local communication are depicted in Fig. 4. An oval on each figure shows the set of division ratios which improves the latency of the proposed network over the conventional one for each percentage of the local communications.

In this case, BDNoC $[(x,y),1]$ does not improve the network performance when the local (non-local) communications are less than 40% (higher than 60%). Dividing the physical channel into two layers when the non-local communications are high, increases the number of flits worsening the traffic congestion in the intermediate routers of a network. For the local communications of 40, 50, and 60 percent, some improvement on the average network latencies may be achieved for the BDNoC architecture. In the case of BDNoC $[(x,y),2]$, even for 40% and 50% local communications, no performance improvement was achieved. This is explained by noting that, in BDNoC $[(x,y),2]$ compared to BDNoC $[(x,y),1]$, both local and non-local packets may pass through more hops. The results show that only halving the width equally between the two layers (*i.e.*, (32,32)) provides us with some latency improvement in the case of 60% local communication.

128-bit links -Now, let us consider the NoC with the physical channel width of 128 bits. Fig. 5 shows the comparison of the average latencies between BDNoC and conventional NoC when the local communications is one hop. Similar results for at most two hops are also depicted in Fig. 6. The optimum division ratios are specified by an oval on each figure.

For instance, the results for BDNoC $[(x,y),1]$ show that for the locality rate of 50% the division ratios of (40,88), (48,80), and (64,64) yield better performances compared to those of the conventional one. 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 the non-local communications include packets traversing more hops than those in the local communications. For this ratio, since the layer *B* width is not large, more cycles are required for these packets to reach the destination. Therefore, one should assign a larger part of the channel width to the non-local communications. Also, note that as is expected, when the share of the local communications increases, a better performance is achieved when larger widths are assigned for the layer *A* (the oval moves to the right). As the results show, for the channel width of 128 bits, for most cases, the BDNoC architectures outperform the conventional one.

If the locality rate varies during the runtime within a limited range, by setting the width of the layer *A* and *B*, BDNoC can still outperform the conventional one. For this purpose, we should determine the range of traffic ratio variation and select the link division ratio which is common among the ratios specified by the ovals shown in Figs. 5 and 6. For example, by choosing the width division of (40,88), BDNoC can make improvement for the locality ratio changing from 30% to 50%. Also, BDNoC $[(64,64),2]$ outperforms the conventional NoC when the local traffic varies from 50% to 70% during the runtime. Hence, although we do not necessarily obtain the maximum improvement, BDNoC is still beneficial as underlying communication architecture for varying local traffic ratios.

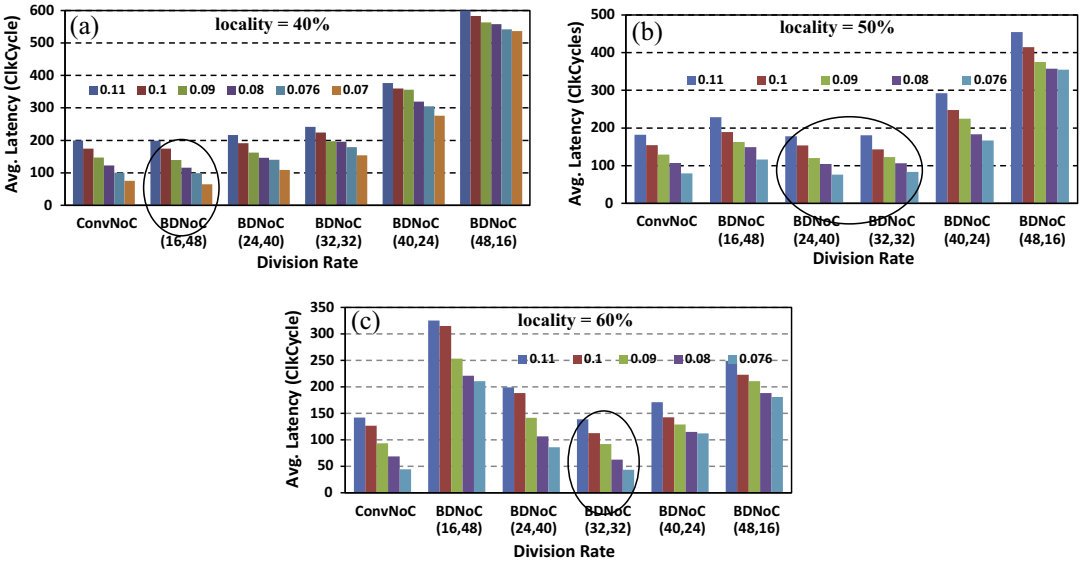

Fig. 3. Average network latencies of BDNoC $[(x,y),1]$ and Conventional NoC with 64-bit links and 5×5 mesh topology under different non-uniform traffic distributions with varying packet injection rate (packet/cycle).

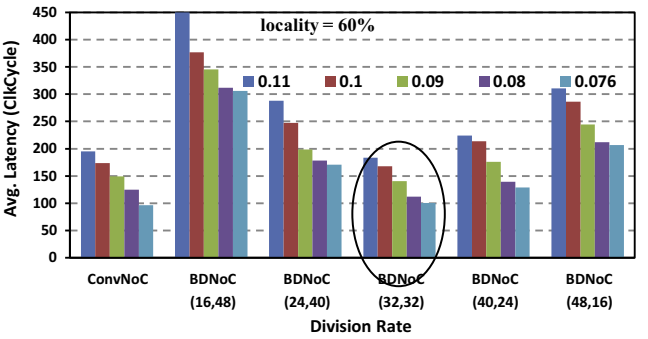

Fig. 4. Average network latencies of BDNoC $[(x,y),2]$ and Conventional NoC with 64-bit links and 5×5 mesh topology under non-uniform traffic distribution with 60% local traffic with varying packet injection rate (packet/cycle).

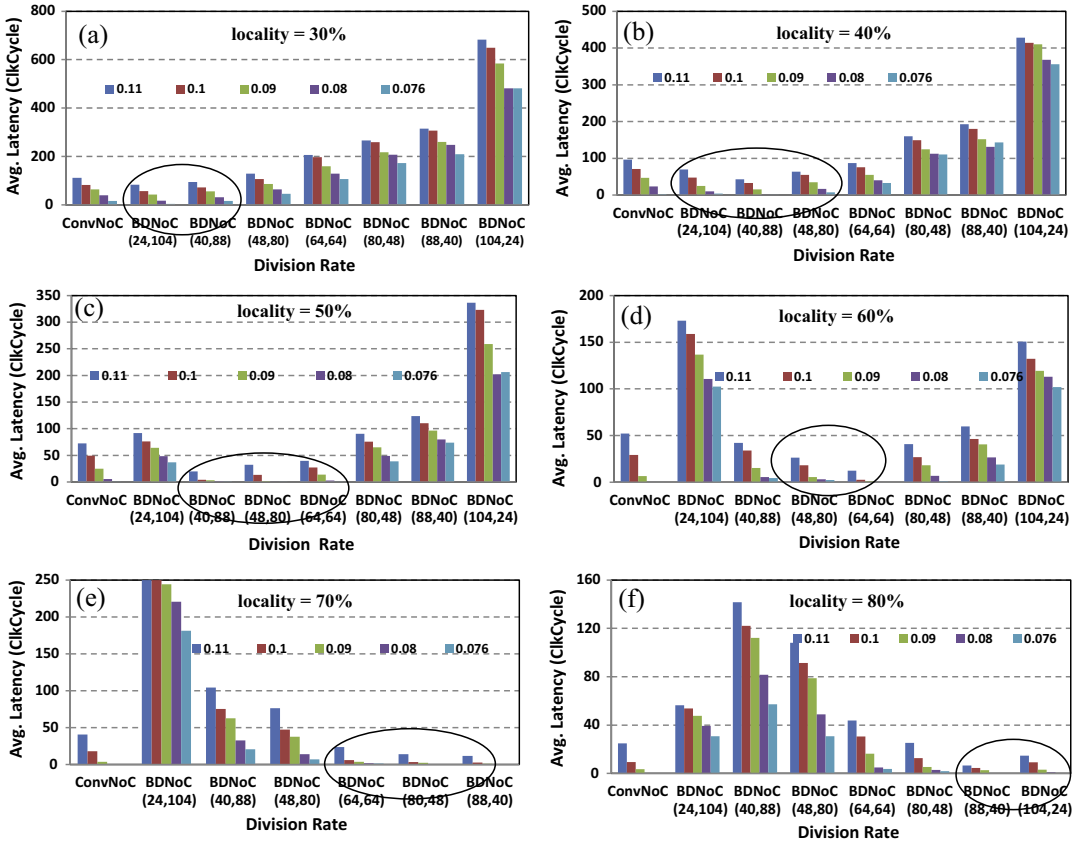

Fig. 5. Average network latencies of BDNoC $[(x,y),1]$ and conventional NoC with 128-bit links and 5×5 mesh topology under different non-uniform traffic distributions with varying PIR (packet/cycle).

For each local communication rate, the results related to division ratios at which BDNoC $[(x,y),1]$ and BDNoC $[(x,y),2]$ outperform the conventional one are individually studied in Figs. 7 and 8, respectively. It is observed that the conventional architecture with over 30% local traffic saturates at a smaller PIR than the proposed one.

For example, in Fig. 7(c), for the communication locality of 50%, the conventional NoC with 128-bit links saturates at the PIR of 0.076 while the BDNoC $[(40,88),1]$ saturates at PIR of 0.1. Also, Fig. 7(c) shows that at higher PIRs, (40,88) leads to a 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 pass through more hops increases more compared to the local ones. To resolve this issue, one should increase non-local link width.

Now, let us consider the case of 50% local communications defined based on two-hop distance. As shown in Fig. 8(c), conventional NoC saturates at PIR of 0.07 while BDNoC $[(48,80),2]$ saturates at PIR of 0.08. Since local and non-local packets

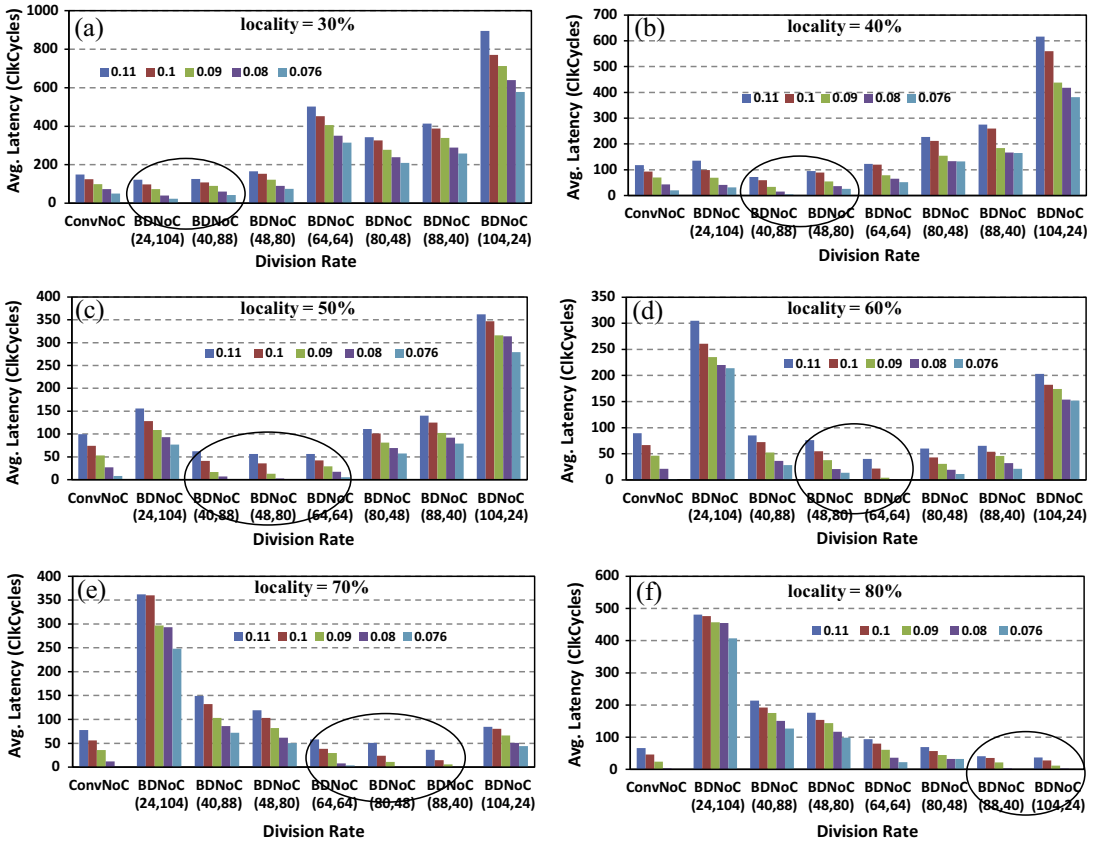

Fig. 6. Average network latencies of BDNoC [(x,y),2] and conventional NoC with 128-bit links and 5 × 5 mesh topology under different non-uniform traffic distributions with varying PIR (packet/cycle).

traverse more hops in this case, the saturation points have been occurred at lower PIRs compared to those in Fig. 7(c). Finally, note that at locality rate ranging from 30% to 80%, the proposed architecture for 128-bit channels provides, on the average, 56% improvement in term of the average network latency compared to that of the conventional one.

In order to compare the average network latencies of the local and non-local packets individually in conventional NoC and BDNoC, their corresponding average network latencies were studied in both architectures. The results for local traffics of 30%, 50%, and 60% defined based on one hop are depicted in Fig. 9. The figures indicate that the improvement in the overall average network latency of the packets in the case of BDNoC originates from the latency reductions of both local and non-local packets.

5.3. Determination of local distance for NED and Uniform traffic patterns

In this section, the performance of the proposed architecture using Uniform and NED traffic patterns is studied. For these traffic profiles, the number of hops used for defining the local communications determines the efficiency of the proposed BDNoC. The study is performed for different width division ratios and mesh sizes. In the 3 × 3 mesh, for the NED traffic profile, one (two) hop local communications, on average, form 50% (80%) of the total communications. In the 5 × 5 mesh, for the NED and Uniform traffic profiles, one hop local communications are on average about 30% and 12% of the local traffics, respectively. By increasing the distance to at most two hops, they become about 60% and 34%, respectively. For at most 3 hops between the source and destination nodes, the local communications comprise 85% and 57%, respectively.

Figs. 10 and 11 provides the average network latency of a 3 × 3 mesh under the NED traffic distribution for channel widths of 64 bits and 128 bits, respectively. As is evident from the figures, in the router with 64-bit channels, BDNoC [(32,32),1], and in the router with 128-bit channels, BDNoC [(64,64),1], BDNoC [(40,88),1], and BDNoC [(104,24),2] provide lower latencies compared to those of the conventional NoC. As was mentioned earlier in Section V.B, for 64-bit channels, since dividing links in general increases the number of flits more compared to the case of 128-bit channels, in the case of BDNoC [(48,16),2], dividing the width into two layers, considerably decreases the performance of the network. The results, also, suggest that for both channel widths (*i.e.*, 64 bits and 128 bits), defining the local communication based on one hop provides lower network latency and higher saturation point. Note that defining the local communication based on two hops

Please cite this article in press as: Akhlaghi V et al. An efficient network on-chip architecture based on isolating local and non-local communications. Comput Electr Eng (2014), http://dx.doi.org/10.1016/j.compeleceng.2014.12.002

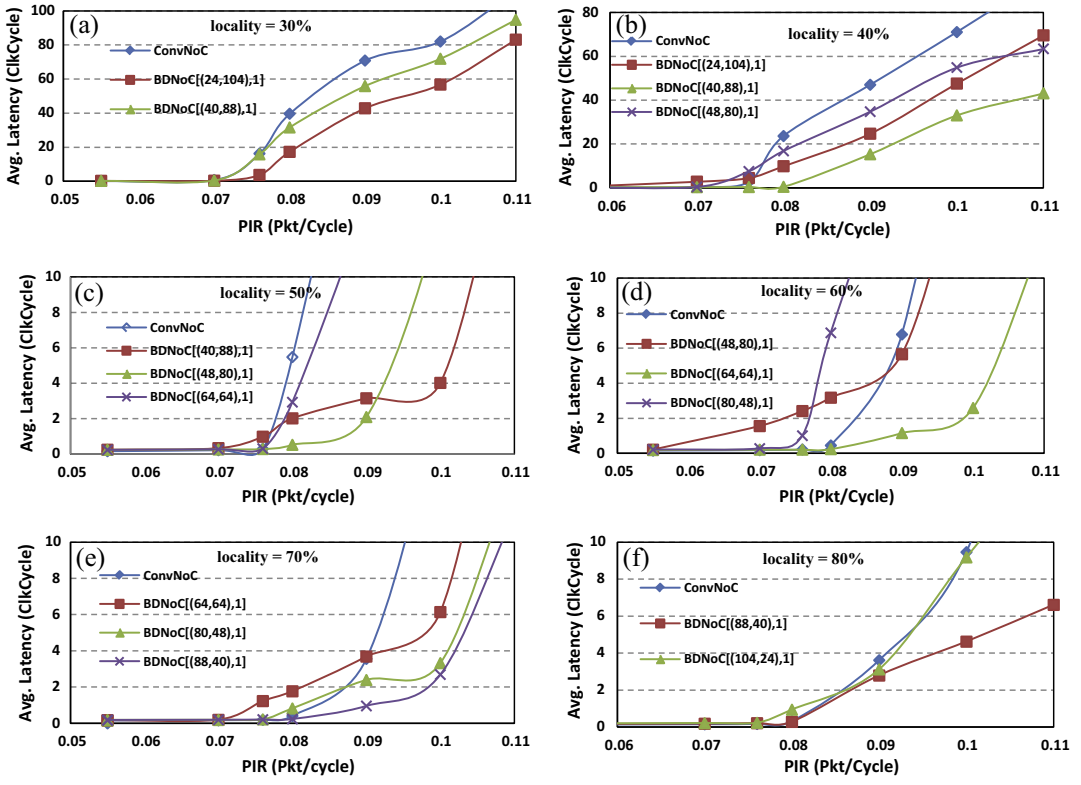

Fig. 7. Average network latencies versus PIR for 5 × 5 mesh under different non-uniform traffic distributions for one hop local communication with 128-bit links.

(one hop) causes that 80% (50%) of all the packets become local ones. The high volume of local traffic adversely affect the efficacy of the proposed architecture in the case of BDNoC [(*x*,*y*),2].

Similar results for the mesh size of 5 × 5 are shown in Figs. 12 and 13, for the channel widths of 64 and 128 bits, respectively. The results indicate that BDNoC [(16,48),1], BDNoC [(32,32),2], BDNoC [(24,104),1], BDNoC [(40,88),1], BDNoC [(64,64),2], and BDNoC [(80,48),2] outperform the conventional NoC. Additionally, the figures reveal that to obtain considerable performance gains, the local communications should be defined based on two hops. This behavior is on contrary to that of the 3 × 3 mesh size. Here, using two hops causes only 60% of the whole communications become local ones.

It should be noted that the reason of using mesh of size 3 × 3 and 5 × 5 is to have different communication patterns for evaluating the efficacy of the proposed BDNoC approach (without being concerned about the actual mesh sizes). For example, by defining local traffic based on a single hop communication, the percentage of local communications changes from 50% (3 × 3) to 30% (5 × 5) under the NED traffic pattern. For these two meshes, non-local packets traverse different numbers of hops affecting the network latency.

Next, to study the effect of the traffic profile on the efficiency of the proposed architecture, the results of the average latency of the 5 × 5 mesh for the Uniform traffic profile were obtained. The results for 64-bit and 128-bit channels are shown in Figs. 14 and 15, respectively. It is observed that BDNoC [(32,32),3], BDNoC [(16,48),2], BDNoC [(64,64),3] and BDNoC [(24,104),2] provide lower latencies than those of the conventional NoC, whereas BDNoC [(16,48),1] and BDNoC [(24,104),1] offer no benefit. The reason is that, by defining the local communications only for one hop packet transmission, gives rise to a non-local communication of approximately 88% which must use the layer *B*. This large volume of traffic leads to the deterioration of the average latency compared to that of the conventional one. Also, the results suggest that BDNoC [(*x*,*x*),3] leads to a higher performance than the others, mainly because, in this case, the local communications form 57% of the total traffics balancing the workloads on the two layers.

Based on the above discussion, it can be concluded that for the NED and Uniform traffic patterns, including around 55.5% of traffic in the local traffic, one can achieve significant improvement (*i.e.*, 50%, on average) in terms of average network latency if the proposed network architecture is used.

5.4. Real application

To study the efficacy of the proposed technique under a real application, BDNoC for the MPEG-4 application has been evaluated. The IP elements of the application were obtained from [26] and mapped on 3 × 4 mesh network using Nearest

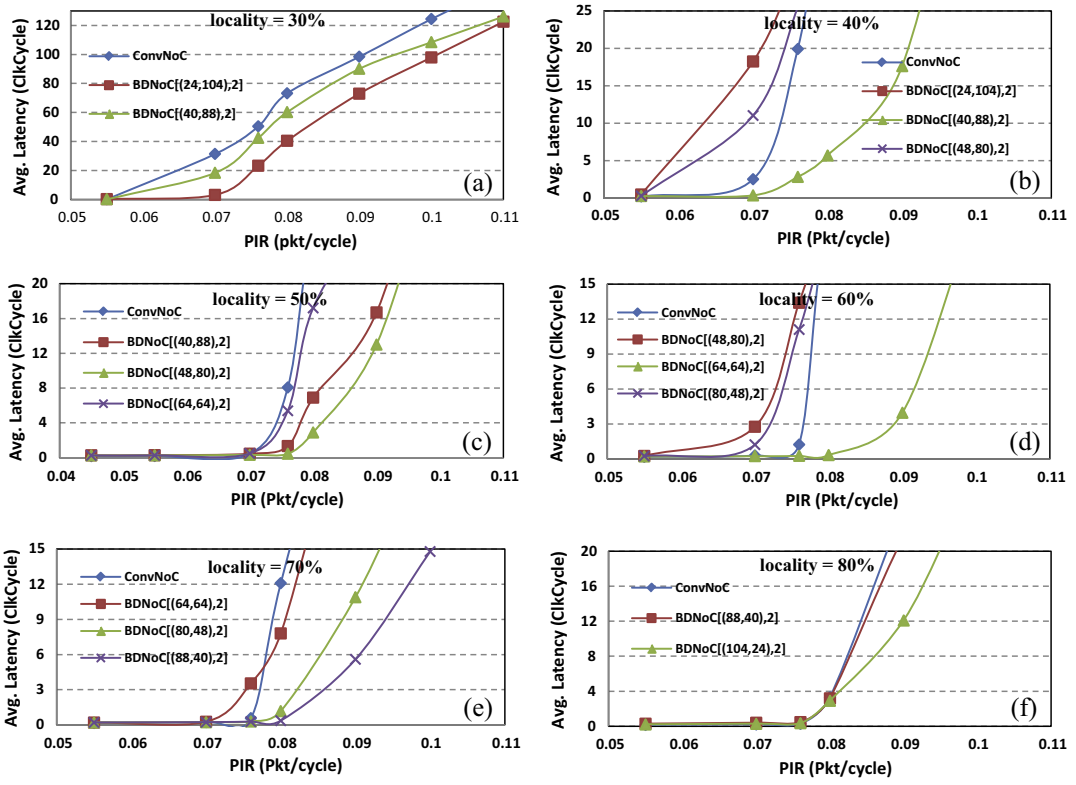

Fig. 8. Average network latencies versus PIR for 5 × 5 mesh under different non-uniform traffic distributions for two hop local communication with 128-bit links.

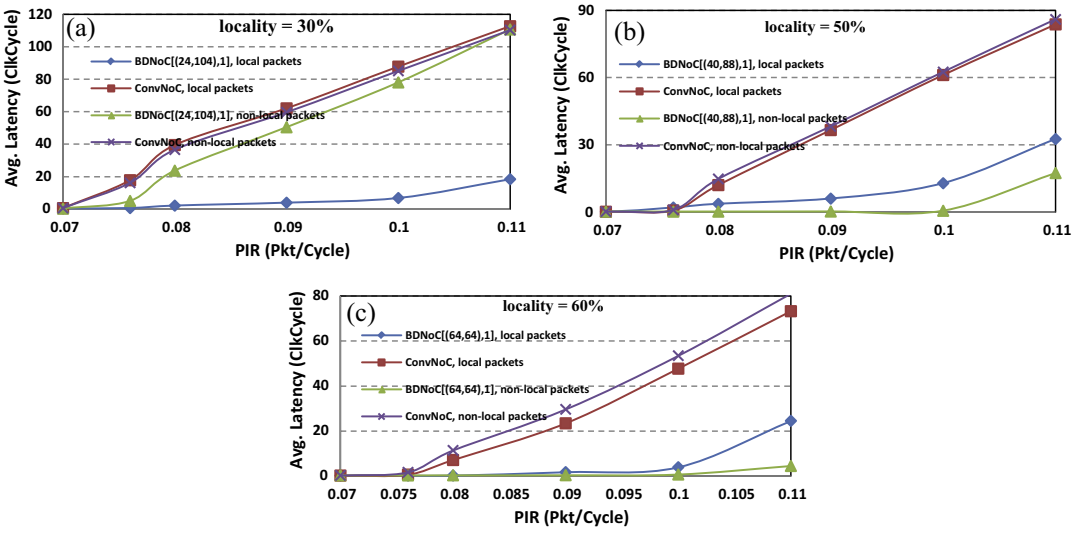

Fig. 9. Average network latencies of local and non-local packets in BDNoC [(*x*, *y*), 1] and conventional NoC with 128-bit links and 5 × 5 mesh topology under different non-uniform traffic distributions with varying PIR (packet/cycle).

Neighbor (NN) and random mapping techniques. The comparison of the average network latencies when BDNoC and conventional NoC (ConvNoC) are used is illustrated in Fig. 16. In this study, the results of the previous section to determine the better sizes for each layer in the BDNoC based on the amount of local traffic was used. For the NN mapping case, the local distance was considered to be a single hop communication making the average locality of the network to about 80% and hence BDNoC [(104, 24), 1] was used. In the case of NN mapping, the BDNoC outperforms the conventional NoC by about 25%. In the case of the random mapping, assuming the local distance to be a single hop communication leads to about

Please cite this article in press as: Akhlaghi V et al. An efficient network on-chip architecture based on isolating local and non-local communications. Comput Electr Eng (2014), http://dx.doi.org/10.1016/j.compeleceng.2014.12.002

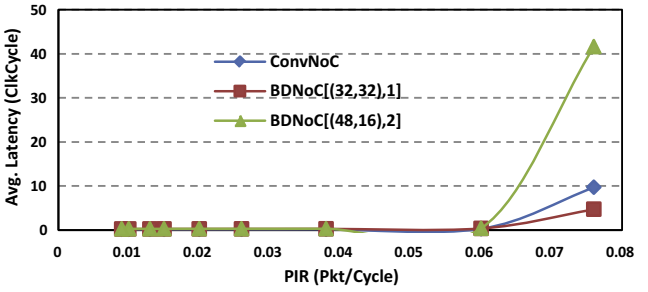

Fig. 10. Average network latencies versus PIR of a 3 × 3 mesh with 64-bit links under NED traffic distribution.

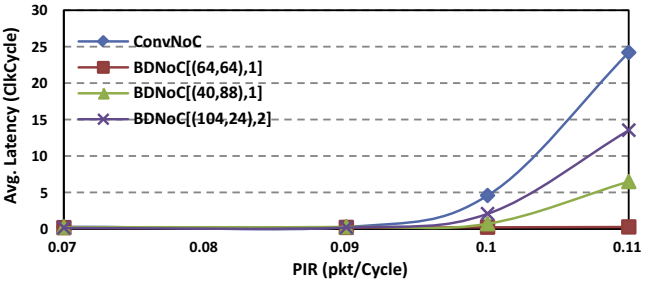

Fig. 11. Average network latencies versus PIR of a 3 × 3 mesh with 128-bit under NED traffic distribution.

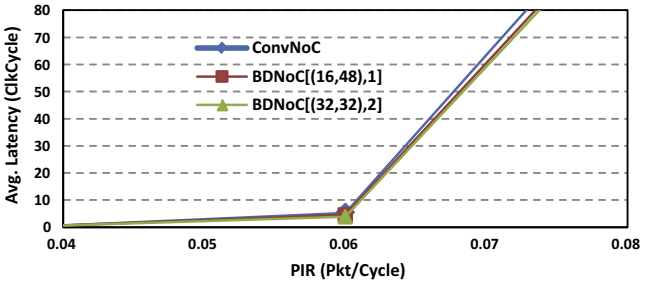

Fig. 12. Average network latencies versus PIR of a 5 × 5 mesh with 64-bit links under NED traffic distribution.

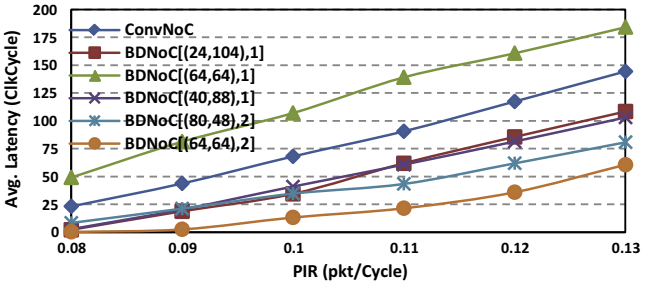

Fig. 13. Average network latencies versus PIR of a 5 × 5 mesh with 128-bit links under NED traffic distribution.

10% local communication. For this case, BDNoC [(24,104),1] was used. Since the percentage of the local traffic is small, BDNoC performs slightly worse than ConvNoC.

As discussed before, the definition of the local distance affects the efficacy of the proposed technique. For the case of random mapping in the MPEG-4 application, local packets can be defined as the packets whose source and destination nodes are away by at most two hops. This leads to, on average, about 50% of local traffics. For this definition of locality, the use of 48 bits for the layer *A* provides us with about 10% improvement in the average network latency.

Please cite this article in press as: Akhlaghi V et al. An efficient network on-chip architecture based on isolating local and non-local communications. Comput Electr Eng (2014), http://dx.doi.org/10.1016/j.compeleceng.2014.12.002

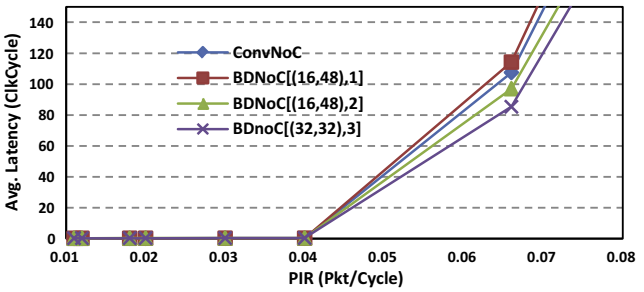

Fig. 14. Average network latencies versus PIR of a 5 × 5 mesh with 64-bit links under Uniform traffic distribution.

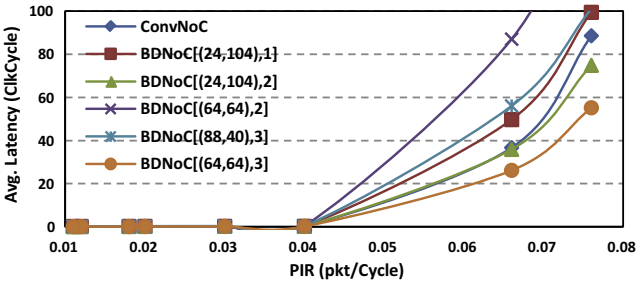

Fig. 15. Average network latencies versus PIR of a 5 × 5 mesh with 128-bit links under Uniform traffic distribution.

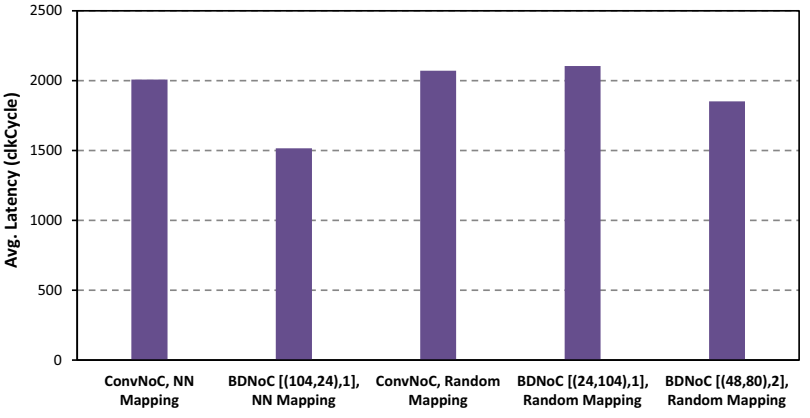

Fig. 16. Average network latencies of BDNoC and conventional NoC (ConvNoC) for MPEG-4 when NN and random mapping techniques are used.

5.5. BDNOC compared to the virtual channel technique

To show the effectiveness of the BDNoC, this architecture was compared with a NoC which has two virtual channels (VCs) of VC0 and VC1. In this network architecture, which is also optimized for the local communication, VC0 has a higher priority. The network was implemented using VHDL and simulated for different non-uniform traffic patterns (*i.e.*, 50%, 60%, 70%, and 80%). It was assumed that the local packets had higher priorities compared to the non-local ones, and hence, VC0 is devoted to the local packets. Also, it was assumed that the channel width was 128-bits, the mesh size was 5 × 5, and the local communication was defined based on one hop distance. The results which are shown in Fig. 17 indicate that BDNoC provides lower average latencies compared to the NoC with VC (VCNoC) for all packet injection rates.

5.6. Energy and energy-delay product

The energy and Energy-Delay Product (EDP) of the proposed and conventional network architectures are studied in this part. For this purpose, a 45 nm standard CMOS Library [25,27] has been used to synthesize the VHDL models of these two NoC architectures. In order to compute the power consumptions of the implemented NoCs, the generated VCD files obtained

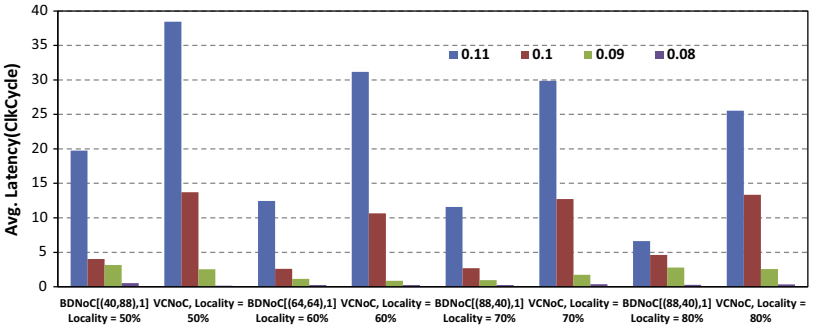

Fig. 17. Average network latencies of BDNoC and NoC with virtual channels (VCNoC) under non-uniform traffic distributions with different locality rates and PIR (packet/cycle). The mesh size was assumed to be 5×5 with 128-bit as the link width. The local communication was defined based on one hop distance.

Table 2
Power consumption of each router.

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

from simulating the design with Modelsim were fed into the Synopsys PrimePower tool. The results for the average power dissipations of the two routers are presented in Table 2. In addition, we have given the figure for each layer of BDNoC. Because BDNoC has two network layers for the local and non-local communications, it dissipates more power than that of the conventional one. To calculate the energy consumption of a packet transmitted through the routers, the average network latency of the conventional architecture has been multiplied by its router power. Note that in this work, only the power of the router was considered and the powers of the wires were not included. In the case of BDNoC, the average network latency of each layer (*A* and *B*) along with the corresponding power consumption were used. Therefore,

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

where α indicates the percentage of the local communications, P_A ($Lat_{Avg,A}$) and P_B ($Lat_{Avg,B}$) are the power consumptions (the average network latencies) of the layers *A* and *B*, respectively.

EDP is obtained by multiplying the energy per packet by the average network latency in the case of the conventional NoC. For the case of BDNoC, the EDP is obtained from:

$$EDP = \alpha \cdot P_A \cdot Lat^2_{Avg,A} + (1-\alpha) \cdot P_B \cdot Lat^2_{Avg,B} \quad (2)$$

The energies and EDP values of the BDNoC $[(x,y),1]$ and the conventional NoCs for different rates of local communications and packet injection rates are calculated and provided in Figs. 18 and 19, respectively. Here, a channel width of 128 bits among the switches was assumed. In calculating the energy and EDP in BDNoC, for each locality rate, the width division ratio

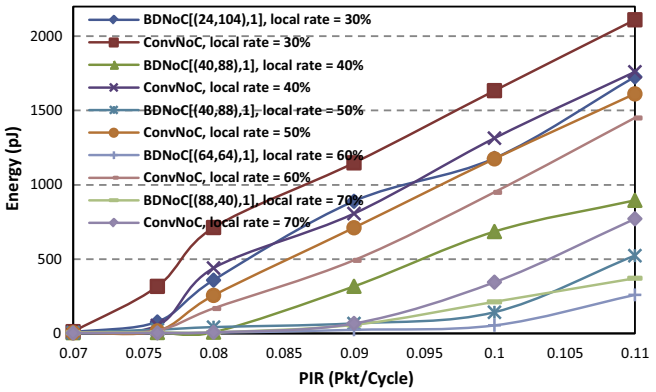

Fig. 18. Energy consumption comparison between BDNoC $[(x,y),1]$ and ConvNoC for different rates of communication locality.

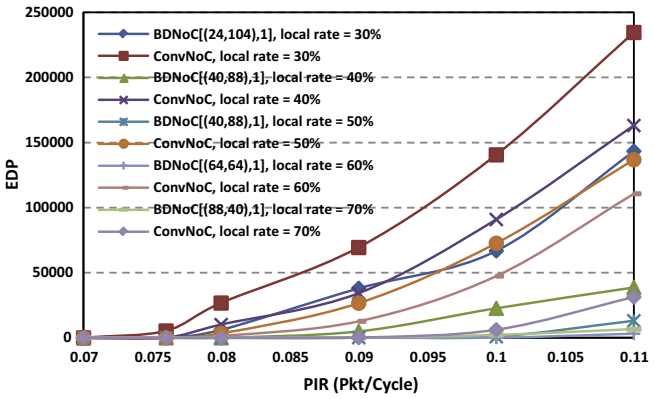

Fig. 19. EDP comparison between BDNoC $[(x,y),1]$ and ConvNoC for different rates of communication locality.

resulting in the optimum performance improvement was chosen. In general, when the packet injection rate exceeded 0.08 packets per cycle, the proposed architecture revealed a superior performance compared to that of the conventional one by about 60%, and 70% in terms of energy consumption and EDP, respectively. Particularly, the local communication rates of 50% and 60%, led to significant improvements because, in these cases, the traffic distributions over the two network layers *A* and *B* are more balanced compared to other locality rates.

6. Conclusion

In this paper, a locality-aware on-chip interconnect architecture was introduced. In the architecture, the local and non-local communications were carried out through separate network resources. In this approach, the channel width was divided based on the ratio of local traffic to non-local one. To determine the efficacy of the proposed NoC architecture, its performance was compared with that of the conventional one for different percentages of local communications, mesh sizes, link widths, and synthetic traffic profiles. Also, different definitions of local packets based on the maximum number of hops between the source and destination nodes were considered in the study. The results showed that for the link width of 128 bits and local communications of more than 30%, the proposed network architecture outperformed the conventional one, on average, by 56%, 60%, and 70%, in terms of the average network latency, energy consumption, and EDP, respectively. In the case of the link with 64 bit, the performance of the proposed network was better than the conventional one for one hop local communications of 40%, 50% and 60% and two hop local communication of 60%, by 3% and 5%, respectively, in terms of the average network latency. Finally, the performance of the proposed NoC for different definitions of local communications under the NED and Uniform traffic patterns was investigated. The results indicated that defining local communications such that it includes between approximately 50% and 60% of all communications, could lead to an improvement of about 50% in average network latency.

Acknowledgment

VA, MK and AAK acknowledge the financial support by the Iranian National Science Foundation (INSF).

References

[1] Taylor MB, Kim J, Miller J, Wentzlaff D, Ghodrat F, Greenwald B, et al. A 16-issue multiple-program counter microprocessor with point-to-point scalar operand network. In: Proceedings of international solid-state circuits conference, vol. 1. 2003. p. 170–71.
[2] Bocchi M, De Dominicis M, Mucci C, Deledda A, Campi F, Lodi A, et al. Design and implementation of a reconfigurable heterogeneous multiprocessor SoC. In: Proceedings of custom integrated circuits conference; 2006. p. 93–6.
[3] Marculescu R, Bogdan P. The chip is the network: toward a science of network-on-chip design. Found Trends Electron Des Autom 2009;2(4):371–461.
[4] Benini L, De Micheli G. Networks on chip: a new SoC paradigm. IEEE J Comput 2002;35(1).
[5] Carvalho E, Calazans N, Moraes F. Heuristics for dynamic task mapping in NoC-based heterogeneous MPSoCs. In: Proceedings of international workshop on rapid system prototyping; 2006. p. 34–40.
[6] Carvalho E, Moraes F. Congestion-aware task mapping in heterogeneous MPSoCs. In: International symposium on System-on-Chip (SoC); 2008. p. 1–4.
[7] Singh AK, Jigang Wu, Prakash A, Srikanthan T. Efficient heuristics for minimizing communication overhead in NoC-based heterogeneous MPSoC platforms. In: Proceedings of international symposium on rapid system prototyping; 2009. p. 55–60.
[8] Afzali-Kusha A, Pedram M, Rahmani AM. NED: a novel synthetic traffic pattern for power/performance analysis of network-on-chips using negative exponential distribution. J Low Power Electron (Am Sci Publishers) 2009;5:396–405.
[9] Dally WJ, Virtual-channel flow control. In: Proceedings of ACM/IEEE International Symposium on Computer Architecture (ISCA); 1990. p. 60–8.
[10] Bjerregaard T, Mahadevan S. A survey of research and practices of network-on-chip. ACM J Comput Surv 2006;38(1):1–51.
[11] Banerjee N, Vellanki P, Chatha KS. A power and performance model for network-on chip architectures. In: Proceedings of ACM/IEEE Design, Automation and Test in Europe Conference (DATE), vol. 2. 2004. p. 1250–55.

Please cite this article in press as: Akhlaghi V et al. An efficient network on-chip architecture based on isolating local and non-local communications. Comput Electr Eng (2014), http://dx.doi.org/10.1016/j.compeleceng.2014.12.002

[12] Mello A, Tedesco L, Calazans N, Moraes F. Virtual channels in networks-on-chip: implementation and evaluation on hermes NoC. In: Proceedings of ACM/IEEE symposium on integrated circuits and systems design; 2005. p. 178–83.
[13] Yoon YJ, Concer N, Petracca M, Carloni L. Virtual channels vs multiple physical networks: a comparative analysis. In: Proceedings of ACM/IEEE Design Automation Conference (DAC); 2010. p. 162–65.
[14] Gilabert F, Gómez ME, Medardoni S, Bertozzi D. Improved utilization of NoC channel bandwidth by switch replication for cost-effective multi-processor systems-on-chip. In: Proceedings of ACM/IEEE international symposium on Networks-on-Chip (NoCS); 2010. p. 165–72.
[15] Kakoee MR, Bertacco V, Benini L, ReliNoC: a reliable network for priority-based on-chip communication. In: Proceedings of ACM/IEEE Design, Automation and Test in Europe Conference (DATE); 2011. p. 1–6.
[16] Das R, Eachempati S, Mishra AK, Narayanan V, Das CR. Design and evaluation of a hierarchical on-chip interconnect for next-generation CMPs. In: Proceedings of international symposium on High Performance Computer Architecture (HPCA '09); 2009. p. 175–86.
[17] Baas B, Yu Z, Meeuwsen M, Sattari O, Apperson R, Work E, et al. AsAP: Afine-grain multi-core platform for DSP applications. IEEE J Micro 2007;27(2):34–45.
[18] Zhiyi Yu, Baas Bevan M. A low-area multi-link interconnect architecture for GALS chip multiprocessors. In: IEEE transactions on Very Large Scale Integration (VLSI) systems, vol. 18. 2010. p. 750–62.
[19] Kumar R, Deshpande H, Choi G, Sprintson A, Gratz P. Bidirectional interconnect design for low latency high bandwidth NoC. In: Proceedings of International Conference on IC Design & Technology (ICICDT'13); 2013. p. 215–18.
[20] Lee J, Nicopoulos C, Lee HG, Kim J. Sharded router: a novel on-chip router architecture employing bandwidth sharding and stealing. In: Elsevier Journal of Parallel Computing (PARCO), vol. 39, 2013. p. 372–88.
[21] Hesse R, Nicholls J, Jerger NE. Fine-grained bandwidth adaptivity in networks-on-chip using bidirectional channels. In: Proceedings of Sixth IEEE/ACM International Symposium on Networks-on-Chip (NoCS); 2012.
[22] Wang L, Kumar P, Yum KH, Kim EJ. An adaptive physical channel regulator for high performance and low power network-on-chip routers. Technical Report; 2010.
[23] Lori A, Robert F, Cathoor F, Shickova A, Verkest D. Spatial Division multiplexing: a novel approach for gauranteed throughput on NoCs. In: Proceedings of Third IEEE/ACM/IFIP international conference on hardware/software codesign and system synthesis, CODES+ISSS '05; 2005. p. 81–6.
[24] Jing M, Yu Z, Zeng X, Zhou L. Time-division-multiplexer based routing algorithm for NoC system. In: Proceedings of IEEE International Symposium on Circuits and Systems (ISCAS'13); 2013. p. 1652–55.
[25] Vangal S, Howard J, Ruhl G, Dighe S, Wilson H, Tschanz J, et al. An 80-Tile Sub-100-W TeraFLOPS Processor in 65-nm CMOS. IEEE J Solid-State Circ JSSC 2008;43(1):29–41.
[26] Saffari M, Lotfi S, Jafarzadeh N, Afzali-Kusha A. Mapping of cores on to diagonal mesh-based network-on-chip. In: Proceedings of Mediterranean Conference on Embedded Computing (MECO'12); 2012. p. 233–38.
[27] FreePDK, AFree OpenAccess 45 nm PDK and Cell Library for university. <http://www.eda.ncsu.edu>.

Vahideh Akhlaghi received the B.S. and M.S. degree in computer engineering from University of Tehran in 2007, and 2011, respectively. Currently, she is pursuing the Ph.D. degree in the department of Computer Science and Engineering at University of California, San Diego (UCSD). Her research interests include variation tolerant low-power design, embedded systems and approximate methods in GPGPUs.

Mehdi Kamal received his M.Sc. and Ph.D. in computer engineering from Sharif University of Technology, and University of Tehran in 2007, and 2013, respectively. Currently, he is a research associate of the Low-Power High-Performance Nanosystem Laboratory at Electrical and Computer Engineering Department, University of Tehran. His research interests include Reliability in nano-scale design, ASIP design, and Low power design.

Ali Afzali-Kusha received his Ph.D. in Electrical Engineering from University of Michigan in 1994. He is currently a Professor of the School of Electrical and Computer Engineering at the University of Tehran and the Director of Low-Power High-Performance Nanosystems Laboratory. He is a senior member of IEEE, and his current research interests include low-power high-performance design methodologies for nanoelectronics era.

Massoud Pedram is the Stephen and Etta Varra Professor of Electrical Engineering at the University of Southern California. He received his Ph.D. in EECS from UC-Berkeley in 1991. He is an IEEE Fellow, an ACM Distinguished Scientist, and past Editor-in-Chiefs of the ACM TODAES and IEEE JETCAS. His research focuses on energy-efficient computing, energy storage systems, and low power electronics.

Please cite this article in press as: Akhlaghi V et al. An efficient network on-chip architecture based on isolating local and non-local communications. Comput Electr Eng (2014), http://dx.doi.org/10.1016/j.compeleceng.2014.12.002