Design of a Low-Latency Router Based on Virtual Output Queuing and Bypass Channels for Wireless Network-on-Chip

Farhad Rad[a], Midia Reshadi[a], and Ahmad Khademzadeh[b]
[a] *Department of Computer Engineering, Science and Research Branch, Islamic Azad University, Tehran, Iran*
[b]*Education and International Scientific Cooperation Department, Iran Telecommunication Research Center, Tehran, Iran*
Corresponding author: Reshadi@srbiau.ac.ir

Abstract: **Wireless network-on-chip (WiNoC) is considered as a novel approach for designing future multi-core systems. In WiNoCs, wireless routers (WRs) utilize high-bandwidth wireless links to reduce the transmission delay between the long distance nodes. When the network traffic loads increase, a large number of packets will be sent into the wired and wireless links and can easily fill FIFO queues at the input ports of routers. In these conditions, head-of-line (HOL) blocking and node congestion may occur and the network communications efficiency tremendously decreases. In this study, a low-latency router was proposed, which employs virtual output queuing (VOQ) and bypass channels to eliminate the congestion of routers and improves network performance. Synthetic traffic patterns were simulated using Noxim simulator and obtained results show that considerable improvement in the latency, total energy consumption and the saturation throughput can be achieved compared to the other WiNoCs.**

Index Terms- wireless network-on-chip, head-of-line blocking, bypassing channels, virtual output queuing, low-latency routers.

I. INTRODUCTION

The demand for High-Performance Computing (HPC) and big data applications has led to the use of multi-core processors in System-on-Chip (SoC). Network-on-Chip (NoC) has become a promising solution for addressing the communication issues in modern SoC design. Although the NoCs have many advantages compared with bus topologies, they face major drawbacks when the network size increases. In this situation, the communication delay and power consumption increased. Several emerging interconnect paradigms, such as three-dimensional (3D) [1], photonic and radio frequency (RF) NoC [2, 3] have been proposed to address the conventional NoCs drawbacks. However, each of these new communication structures has its own disadvantages. For example, RF-based architectures need additional physically

Manuscript received 12-March-2019 and revised 5-June-2019, *P- ISSN: 2322-4088*
Accepted on 18- Sept.-2019 *E- ISSN: 2322-3936*

overlaid transmission lines and power budget requirements, photonic components are sensitive to temperature and not compatible with CMOS technology and in 3D NoCs, failure in vertical links leads to poor yields [4].

In recent years, the concept of wireless network-on-chip (WiNoC) is proposed to compatible with the existing CMOS technology for reducing the communication delay between multi-hop nodes. [5]. In WiNoCs, a wireless router (WR) is a baseline router (BR) that in addition to FIFO input buffers, routing computation logic, virtual channels (VCs) allocator, switch allocator, and crossbar switch is equipped with antenna and transceivers to achieve the wireless interconnections. Wireless links have higher bandwidth and lower delay compared to wired links. Hence, these links are able to transmit data packets between long distance nodes. In heavy traffic loads, if simple FIFO buffer queues used in the input ports of both routers, head-of-line (HOL) blocking occurs. This phenomenon leads to node congestion and thermal hotspot forms in the network. Therefore, the network performance can be reduced. On the other hand, in WiNoCs, routers transfer packets between the short distance and long distance nodes with wired and wireless links. Thus, wired and wireless links, BRs, and WRs are the most important latency parameters in these networks. The latency of a packet over a multi-hop path with wired and wireless links is calculated based on the clock cycle using the following equation:

$$T = T_{BR}h_{Wire} + T_{WR}h_{Wireless} + T_{WiredL}(h_{Wire} - 1) + T_{WierlessL}h_{Wireless} + L/BW_{Wire} + L/BW_{Wireless} \quad (1)$$

In which h_{Wire} and $h_{Wireless}$ are the number of hops in wired and wireless path, T_{BR} and T_{WR} are the delay of the baseline router and the wireless router, $T_{Wiredlink}$ and $T_{WirelessL}$ are the communication delay of the wired and wireless links between two routers, L/BW_{wire} and $L/BW_{Wireless}$ are times required for a packet with length L to cross a wired or wireless link with bandwidth BW. According to Eq. (1), when communication latency and queuing delay in each router minimized, the performance of WiNoC improve. Therefore, the use of low-latency routers can reduce a fraction of the communication latency overhead.

The main contribution of the paper is the design of a low-latency router based on the virtual output queuing (VOQ) and the bypassing channels to improve the network performance of the WiNoCs. In the proposed scheme, the use of multiple VOQs eliminates the HOL problem in the input buffer queues, and bypass channels will reduce pipeline stages. Also, to utilize the advantages of the proposed scheme, an adaptive routing algorithm for transmitting packets at low and high traffic loads is proposed.

The rest of the paper is organized as follows: In Section II, related work is briefly reviewed and discussed. Then, in Section III, the WiNoC topology, proposed routers architectures, virtual channel allocation, and an adaptive routing algorithm are presented. Section IV shows the simulation results. Finally, Section V concludes this paper.

II. RELATED WORK

In recent years, with the advent of a variety of CMOS compatible antennas, the concept of WiNoC has been introduced more seriously to solve scalability and performance bottleneck of the conventional NoCs. The carbon nanotubes (CNTs) antenna and millimeter-wave antennas have been demonstrated for long-range communications on the chip [6, 7]. The achieved bandwidth of millimeter-wave antennas is 10-100 GHz, while in CNTs antenna is around 500 GHz. Design and implement of sub-THz frequency range antenna in a polyamide layer with 20 Gbps data rate is proposed in [8]. Wireless data communication links based on Graphene antennas is proposed in [9]. These antennae can operate at THz frequencies range with higher bandwidth (× 100 Gbps), and higher data rate compared to mm-wave transceivers. A WiNoC architecture based on ultra-wide-band (UWB) technology with the high data rate and the short-range communication has been proposed in [10].

In addition to the above-mentioned wireless technologies, different kinds of the WiNoC architectures have been proposed to demonstrate a considerable improvement of the network performance in comparison to the traditional wire NoC architectures [11-14]. Hybrid architectures that used both wired and wireless links show an excellent trade-off between latency and energy consumption. In the hybrid architecture, the whole network divided into smaller subnets. A WR is located at the center of each subnet and is responsible for directly connecting to another subnet. Using the wireless links between the long distant nodes can reduce the hop count. Consequently, it leads to low latency and high power efficiency in the WiNoCs. However, in heavy traffic loads, the demand for using WRs increases and led to the amount of data exceeds the bandwidth capacity of the wireless links. Hence, congestion can occur. The congestion reduces the efficiency of the communication and causes degradation of the network performance [15]. Therefore, various methods such as congestion-aware WiNoC architectures or congestion-aware routing algorithms have been proposed to alleviate the congestion of WRs and balanced the network load in the WiNoCs [13, 16-18]. Some of these studies combine congestion-aware routing algorithms with application mapping and task migration methods to overcome the congestion of the WiNoCs [13, 19]. These schemes focused on load balancing to solve the congestion problem in the WiNoCs, and they do not need additional hardware resources or increasing the buffer size.

In [20], a flit counter and address resolution unit are proposed to add in the WRs. To determine the congestion status of each WR, these units collect the congestion information and destination address information and then a congestion judgment algorithm set the highest priority to transmission based on this information. This solution can effectively avoid increased congestion at the wireless nodes, but information controlling at wireless nodes will to some extent restrict the utilization efficiency of the wireless resource. Also, the HOL blocking issues is not present in this work.

To eliminate the HOL blocking, VOQ mechanism can be used. Although VOQ is succeeded in the HOL blocking problem, the performance will decline when the traffic load becomes heavy. In [21], for the first

time, an NoC based on multiple VOQ is proposed. The authors suggested that each input port maintains multiple independent queues for each output port. The results of this work in the heavy traffic loads have shown that using the multiple VOQ can to eliminate the HOL blocking and reduce the congestion at the output ports. In [22], a novel WR architecture based on multiple VOQ is proposed. The authors investigated a turning-restrictions communication scheme to alleviate congestion in the wireless nodes and hence improve network performance. Although their proposed WR architecture is useful at the network performance, they do not use the bypassing mechanism to reduce the router pipeline and queuing delay. In a bypassing mechanism, the packet buffering process is not be required at the input ports and however, a significant part of the router pipeline and queuing delay can be removed within routers.

The NoCs based on multi-hop traversal in a single cycle has been introduced in [23, 24]. In [24], although the multi-hop distance is replaced to a single-hop distance, it requires extra wireline and power overhead. Also, this scheme is not compatible with the hierarchical topologies which used the long links. In [23], an improved design of SMART routers has been introduced. This scheme no required many control wires in contrast to SMART and is compatible with all NoC topologies.

In [25], a low-latency router with single-cycle bypassing mechanism has been presented. This proposed router is a 3-stage adaptive VC compatible router which can work in the 3D NoCs or WiNoCs. However, the router must decide at each cycle for passing packets through the bypass or pipeline data path. This decision logic can increase the router latency and the critical path length of the bypassing logic. To solve this problem, a dedicated VC for bypassing paths has been proposed in [25]. Although bypassing with dedicated VC is faster, an extra VC can increase the buffer space in the baseline routers.

In the proposed method of this paper, a bypass path is determined only once for each packet and remains constant during the lifetime of the packet. Also, multiple VOQ used in the input ports can remove the HOL blocking. Finally, simulation results show that by integrating bypass channels and a novel adaptive routing algorithm in VOQ based WiNoCs, the network performance will be improved.

III. THE PROPOSED SCHEME

In this section, we first present the WiNoC topology. Subsequently, we show the details of the proposed BR and WR micro-architectures design including the virtual output queuing and the bypass channels. Finally, an novel adaptive routing algorithm is presented.

A. The WiNoC topology

In this paper, a hybrid WiNoC topology is built on a 10×10 two-dimensional mesh network. At first, the mesh network is divided into four 5 ×5 subnets. In the center of each subnet, a BR is replaced by a WR which equipped with a wireless interface (WI). This component is responsible for transmitting packets between long distance nodes by means of the wireless links. In a subnet, each hop is consists of a router

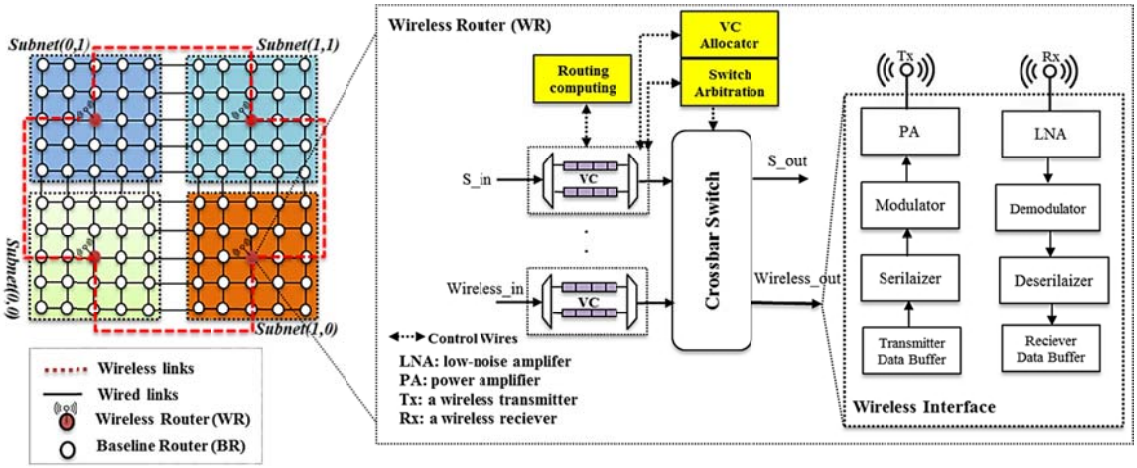

Fig. 1. The network topology with WR architecture.

with the wired or wireless link. The packet transfer requiring one or more clock cycles in each hop. When the network traffic load patterns increase, consequently the hop count increases and the packet latency grows.

The WiNoC topology and the major components of the WRs are illustrated in Fig. 1. A WR includes input buffers with virtual channels (VCs), routing computation logic, virtual channel (VC) allocator, switch arbitration (SA), crossbar, and a wireless interface. The baseline routers are VC compatible routers which use multiple buffer queues to store packets. The architecture is shown in Fig. 1 has several disadvantages as follows. Firstly, the HOL blocking in the input ports may occur when the network traffic loads increase. In the HOL blocking, if the first packet at the beginning of the FIFO buffer queue is blocked then the other queuing packets will not be transmitted and also be blocked. This phenomenon seriously decreases the network performance. Secondly, many clock cycles are required to pass a packet through pipeline stages of the routers. Obviously, the delay of the router's pipeline stages can increase the network latency. Third, the control blocks which are colored with yellow in this figure, lead to a large amount of overhead during the next clock cycles. If these blocks at the pipeline stages are disabling, the network performance improves.

To solve these drawbacks, we proposed the following solutions in this paper:

1) The proposed scheme used multiple VOQ at the input ports to eliminate the HOL blocking under heavy traffic loads. In multiple VOQ, at least two dedicated-single queues are considered in the input ports of the routers for each output port.
2) To reduce the number of pipeline stages on routers, we used a single-cycle bypassing mechanism. The bypass data path can provide shorter paths between routers. In this case, an incoming flit can go directly to the output port without the pipeline stages. In this scheme, only header flit passes

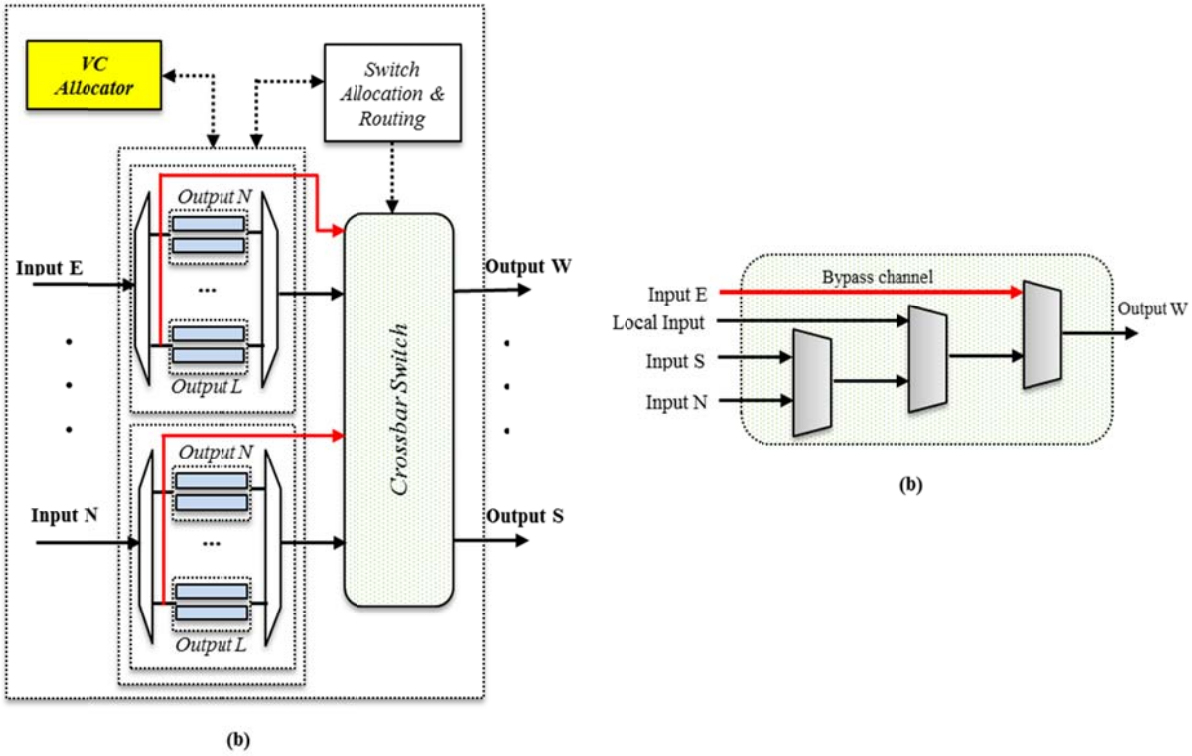

Fig. 2. (a) The proposed BR architecture with multiple VOQ (b) and bypass channel in crossbar switch.

through the router's pipeline and body flits and tail flit pass intermediate routers without buffering process with a single-cycle delay. Its mechanism can achieve better performance than the other methods.

3) A new adaptive routing algorithm has been developed to take advantages of the proposed routers.

B. DESCRIPTION OF THE PROPOSED BASELINE ROUTERS (BRS)

The WiNoC architecture consists of two types of routers; BRs and WRs. The proposed BRs including five ports (east (E), west (W), north (N), south (S), and local injection port (L)), multiple VOQ at the input ports, a routing computation unit, a crossbar switch, a VC allocator (VA) for selection input VCs, and a switch allocation (SA). Fig. 2 illustrates the proposed BR architecture based on multiple VOQ with two VCs for each output port and bypass channels for each input port. The red arrows in Fig. 2(b) represent a low-latency bypass channel that connects the east to west port directly. When a packet takes advantage of the bypass channel skips the crossbar switch procedure and connected to the opposite side in a router (straight bypass). For each incoming packet in each input ports, when the bypass channel is not available, the packet is buffered in the input buffer queues and the general pipeline stages such as routing computation, VC allocation, SA, switch traversal (ST), , and link traversal (LT) are followed to transmit it to the next router. Similar to the work done in [23], the proposed BR in this article uses header flit to

Table I. Pseudocode of VC allocator and switch allocation unit in the proposed baseline router.

Bypass based VC allocator algorithm	**The priority-based switch allocation algorithm**

Input: Incoming packet **Output:** VC number of the input port
for *each packet that arrives from each input port* **do**
if *incoming flit is header flit* **then**
Route incoming flit;
if *(Header flit is a bypass flit)* **then**
Send incoming flit to the output port;(an input VC in the downstream router)
The output port is unreachable;
endif
elseif
Assign a VC number to this packet ;
Update VC number table;
Store the incoming flit in the FIFO buffer;
endif
end for
	Input: Input port request **Output:** The output port
	if *(input port request! = 0)* **then**
	for *each direction of router* ***do***
	if (*Received request is a bypass flit* or *the output is reached through the bypass)* **then**
	Free the output port and make it available to the bypass flit;
	The output port is unreachable;
	Flit traverses without crossbar switch;
	Followed by the link traversal stage;
	elseif *Received request is a non-bypass route* **then**
	the output port number assigned to this input request according to round robin way;
	The output port is unreachable;
	end for
	elseif no input port is authorized;

sets the bypass channels in each router. As long as the bypass channel is maintained progress by a header flit, other flits will be allowed to cross the path without buffering process of the input ports. Therefore, bypass channels reduce the pipeline operation from 4 clock cycles to one clock cycle. On the other hand, to reduce the HOL blocking in the input buffer queues under heavy traffic loads, we used multiple VOQ in each input port. The input ports maintain two independent queues for each output port. In this situation, the packets congestion in one input port will not affect the packet flowing into other output ports. The VC allocator determines that which of the output virtual channel among VC candidates should be assigned to the incoming packet. Upon completion of the VC allocator step, the switch allocation (SA) stage commences. At the end of this stage, the winner flit will occupy an output port and cause this port to be unavailable for all the other packets. In the following, the body flits and tail flit inherited its desired output port. This output port will be free after the tail flit of a packet is transferred. In the proposed BR, the flits of the bypass channel have the highest priority than other flits. The pseudo-code of the VA and SA unit is shown in Table I.

C. Description of the proposed wireless routers (WRs)

In the previous section, a set of bypass channels applied in VOQ compatible baseline routers to provide shorter paths between routers. The proposed wireless router architectures are similar to the BRs but in addition to the mentioned ports has Tx/Rx port. These ports are the wireless interface ports that used to

transmit wireless packets between the long distance nodes. Both proposed routers employ adaptive routing to achieve the shortest route. Although multiple VOQ compatible routers can be used for all baseline routers, for two reasons, its practical implementation in wireless routers are not easy. Firstly, the design of multiple VOQ in the input ports of WRs consumes more hardware resources. In this case, complex control logics are needed to access the shared input VCs. Secondly, WRs are responsible to transmit packets in the short distance as well as the long distance simultaneously. In this situation, many flows compete inside a WR at the same time and so, a WR can be a thermal hotspot in network. Hence, the proposed WR in this study is similar to the proposed WR scheme in [22]. In this scheme, the incoming packet in E, S, W, and N input ports can be only forwarded to three directions (opposite output port, L or WI). Although it scheme is useful for decreasing congestion and increasing performance in WiNoCs, it did not use the bypass channels in BR/WR routers respectively. The bypass channels can lead to creating a single-cycle data path from source to destination nodes. Hence, low-latency communications are achieved. On the other hand, with a set of custom VOQ routers, network performance can be improved by removing the HOL block.

To achieve the mentioned goals, a novel WR router architecture with an adaptive routing algorithm is proposed in as follow. The detail of the proposed router is shown in Fig. 3. The proposed WR router uses multiple VOQ which is customized in each input port. The bypass data path in the WRs is set up by the header flit and the body flits and tail flit pass through the routers without the buffering process. Each input port maintains two independent queues only for each opposite output port. These input queues share the bandwidth of a single physical VOQ channel. The detail of the adaptive routing algorithm scheme is present in subsection *D*. Wormhole flow control mechanism is considered in all routers. The red lines in Fig. 3 are the bypass channels that connect the input ports directly to the output ports. When the header flit is set as bypass flit and the output port is idle, the output port is assigned to the input packets. The packets can bypass the current router with a single-stage pipeline. In this case, the output port is unreachable for other packets of VC queues. The VC allocator determines the VC ID of a packet from the idle VC pool and this selection is not changed at during time. In our proposed scheme, to reduce the router pipeline stages, the VC ID can be allocated for a packet in the routing computation stage. The switch arbitration unit judges between the VC input queues and the output ports. In this unit, the normal flits contending with the bypass flits and finally, only and only one winner will occupy an output port and make it unavailable for all other packets.

In the proposed WR, bypass flits have the highest priority than other flits in the arbitration process. Therefore, they will win the output port and traverse the crossbar switch. A round-robin fashion selects the winner of the output port from multiple requests which have the same priority. The output port will be

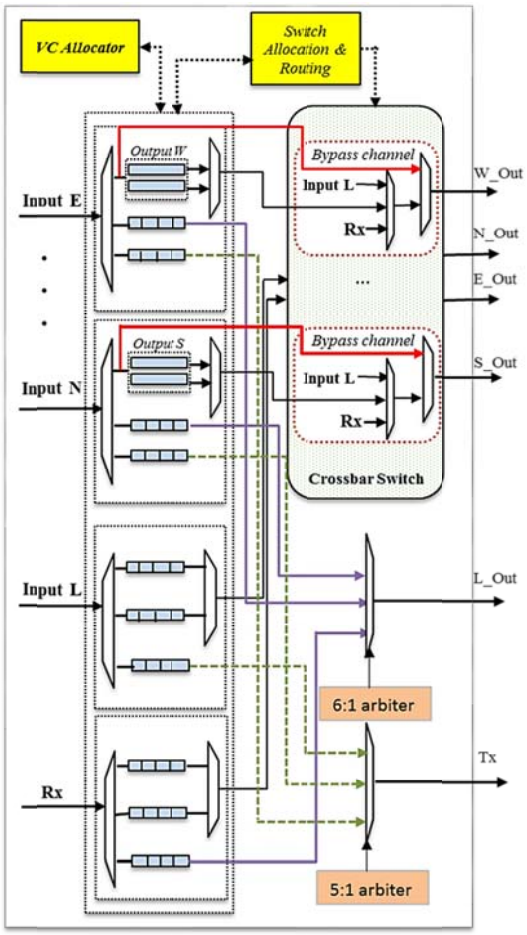

Fig. 3. The proposed WR micro-architecture.

free after the tail flit of a packet is transferred. The algorithms for the input VCs selection (VC Allocation stage) and the SA unit in WRs are similar to the pseudo-code written in Table I.

D. Adaptive Routing algorithm

To reduce the data congestion in WRs due to the internal data flow in each subnet and external data flow between two subnets, an adaptive routing algorithm is adapted to bypass routers. The proposed bypass routers can be using a deterministic routing algorithm, such as XY routing, when the source and destination nodes are in the same subnet or the adjacent subnet, and a long-distance routing algorithm otherwise. In the long-distance routing, using the wireless links, hop count is reduced compared to hop count on the wired network. When a received flit is a bypass flit, the proposed BRs send it without buffering process according to the XY routing and the bypass WRs send it according to the turning-restrictions scheme. Otherwise, it is written to the input buffers and route computation is performed. Once a route is set as a bypass path, an idle VC of the output port is dedicated to the bypass flits. The remaining VCs will be used for other flits. Note that only one virtual channel will be reserved for the bypass flit and make it unavailable for the other flits. In each input port of the routers, we require two VCs (VC0, VC1)

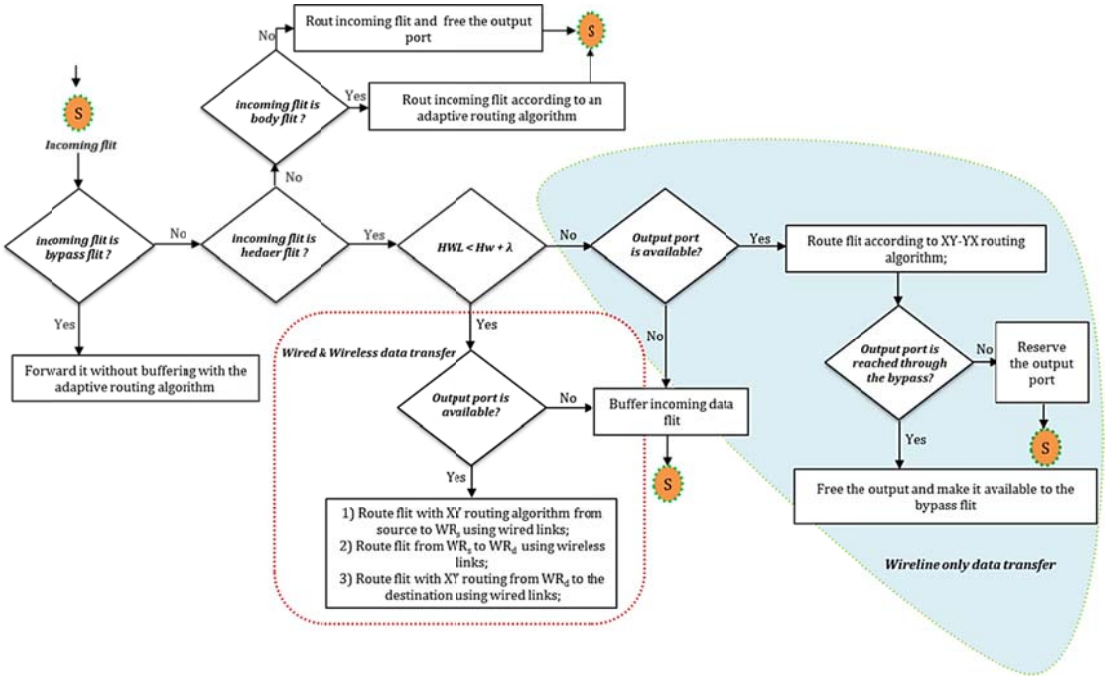

Fig. 4. The proposed adaptive routing algorithm

to prevent routing deadlock. Fig. 4 illustrates all the steps associated with the proposed routing strategy. The parameter λ is used to balance the wired and wireless links utilization. This parameter depends on changing traffic patterns and the network size, so it is experimentally obtained through numerous simulations. In this algorithm, H_{WL} refers to the minimum hop count between the source and destination nodes using both wireless and wired links and H_W is the minimum traveling distance between the source and destination using wired links. When $H_{WL} \le H_W + \lambda$, the non-bypass header flits are allowed to traverse the intermediate routers using the wired and wireless channel. In this case, at the first, the flits are sent using the wired links from the source to nearest wireless router (WR_s) which is located on the same subnet. Then, WR_s sends it using a wireless link to the wireless router (WR_d) that is residing in the destination subnet. Finally, WR_d forwards it to the final destination node using the wired links. All flits in the same subnets or adjacent subnets are transfers through the wired links by a deterministic XY routing algorithm. The pseudo-code of the proposed adaptive routing algorithm is shown in Table II.

IV. SIMULATION RESULTS

In order to evaluate the performance of the proposed routers, a SystemC based Noxim simulator [26] is used. In this paper, Noxim simulator is developed for implementing BRs/WRs and adaptive routing algorithm. The simulation parameters are shown in Table III. The WiNoC topology is considered as a 10 × 10 two-dimensional mesh which is divided into four 5×5 subnets. At the center of each subnet, only one

Table II. Adaptive routing algorithm

Routing Algorithm

Input: Incoming flit, λ **Output:** the output port
Routing (incoming flit);
if *incoming flit is bypass flit* **then**
Forward it without buffering with the adaptive routing algorithm;
endif
if (*incoming flit is header flit*) **then**
if *(src_id & dst_id is on the same subnet or on the adjacent subnet*) **then**
Route flit according to XY routing algorithm;
if *the output is reached through the bypass* **then**
Free the output and make it available to the bypass flit;
endif
elseif
Reserved the output port;
endif
elseif $(H_{WL} < H_w + \lambda)$ **then**
Route flit with XY routing algorithm from source to WR_s using wired links;
Route flit from WR_s to WR_d using wireless links;
Route flit with XY routing from WR_d to the destination using wired links;
elseif
if *incoming flit is body flit* **then**
Rout incoming flit according to an adaptive routing algorithm;
endif
elseif *incoming flit is tail flit* **then**
Rout incoming flit and free the output port;
endif
endif

Table III. Simulation parameters

Parameter	Value
Mesh network size	10×10, 4 subnets
Number of WR	4
Number of virtual-channels per port	4
Packet length [flit]	8
λ	4
Flit size[bit]	32
BR input buffer depth [flit]	4
WR input buffer depth [flit]	8
WR antenna buffer size [flit]	16
Wireless data rate [Gbps]	32
Switching technique	Wormhole
Media access control	Token-based
Packet injection rate (PIR)	0.1
Simulation_time cycles	10000
Stats_warm_up_time cycles	1000

Table IV. The synthetic traffics.

Traffic	Describe
Random	The source nodes generate packets to random destination nodes with uniform probability.
Transpose1 Transpose2	The source nodes generate packets to specific destination nodes.
Bit-reversal	The source nodes generate packets only whose address is the inverse of the sender's address.

WR is used. We evaluate the performance of the proposed schemes compared with a conventional 2D-Mesh without WRs, a VC based WiNoC, and the proposed scheme in [22]. In order to ensure the accuracy of the experimental results, network size, traffic pattern and other experimental parameters are the same. The wireless routers have four virtual channel per port (two for the opposite output port, one for the local port, and one for wireless interface Tx/Rx) and have a buffer depth of 8 flits. The baseline routers have the same parameters including five bi-directional ports, 32-bit data width, and four virtual channels per port with 4-flit buffer size (two 2-flit VCs for each direction). The synthetic traffic patterns have been considered to evaluate the performance of the proposed routers. For the synthetic traffic patterns, we have applied four synthetic traffic patterns, Random, Bit-reversal, Transpose1, and Transpose2. The description of the synthetic traffic patterns is shown in Table IV.

The packet injection rate (pir) is the average number of packets injected per cycle. A pir of 0.1 packets/cycle/node means each node sends a packet (8 flits) every 10 cycles. Also, in the network saturation throughput the wired links are saturated completely and the wireless links are at their maximum utility. The average latency and throughput saturation are used as criteria for evaluating the performance of WiNoCs. Fig. 5(a) shows the average packet latency under the random traffic pattern. In the beginning, the average latency of all schemes is approximately similar except for the 2D-Mesh. The conventional 2D-Mesh latency increases rapidly due to multi-hops data communications between nodes. In other WiNoCs, due to the use of wireless links, the distance between nodes located in different subnets can reduce and thus the network's saturation point has been increased to a large extent. When the injection rate is 0.8 flit /node /cycle, the average latency of the proposed scheme is the lower than of the two other schemes. For the proposed scheme, the network's saturation point is 0.9 flit/node/cycles which are improved by about 12% compared with the VC-based scheme. The network's saturation point of [22] and our scheme is similar but at this point, the average latency of the proposed scheme is by about 33% lower than [22]. This improvement is due to using the bypass channels which can reduce the pipeline stages from four cycles to only one cycle.

Fig. 5(b-d) shows the average latency under two transpose traffic patterns and Bit-reversal traffic pattern respectively. When the traffic load becomes heavier, the average latency of VC based WiNoC under transpose traffics is increase due to the HOL blocking at the input ports. For the scheme proposed in [22], although the congestion has increased in wireless routers due to increasing the packet injection

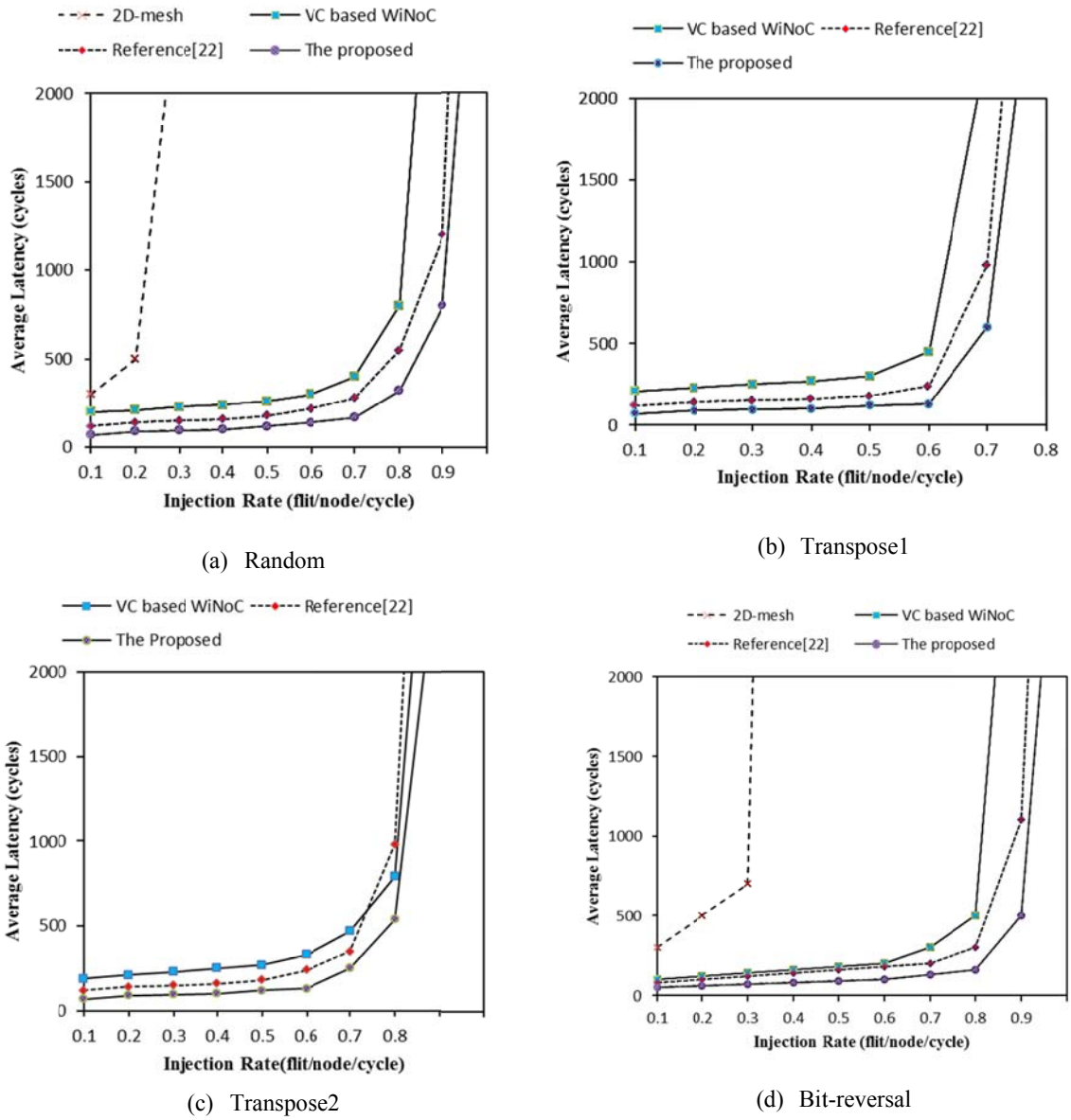

Fig. 5. Average latency comparison of different schemes under different traffic patterns.

rate, the turning-restriction scheme reduces WRs traffic load pressure and leads to the average latency improved. However, in the proposed scheme, the network's saturation point under transpose1 and Bit-reversal traffic loads is improved by about 16%, and 12%, compared with VC based scheme respectively. This is because the wireless links reduce the number of hops in the traversing between the source and destination nodes and the bypass channels reducing the routers pipeline stages. So, the average latency is improved. The advantage of using multiple VOQ in the proposed routers is that congestion at one input port could not affect other input ports. Hence, when a port of the router is congested, flits could still access the output ports. Our proposed method based on the bypass channel forward the packets into the

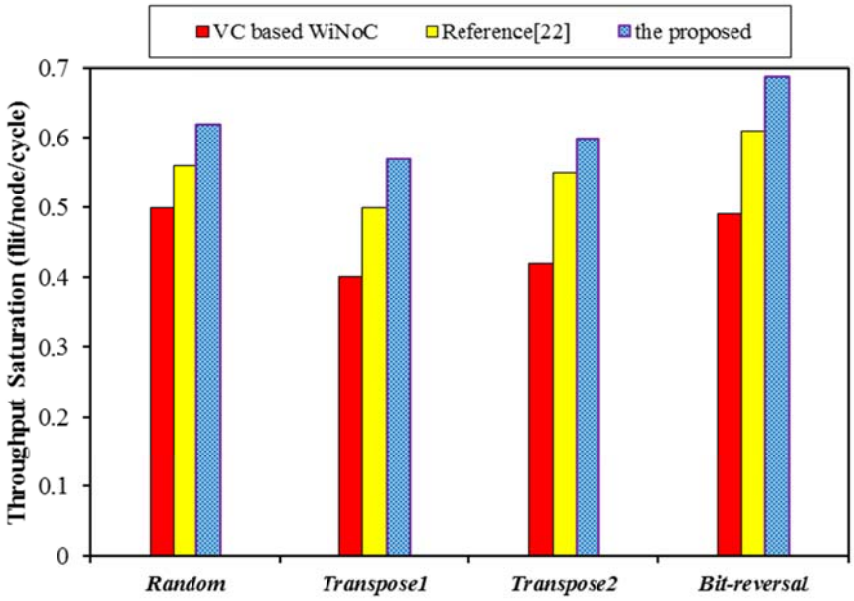

Fig. 6. Throughput saturation under different traffic patterns.

intermediate routers and however have a better traffic load compared to the applied method in [22]. When the network injection rate reached to the maximum value, the average latency of the proposed scheme under transpose1, tarnspose2, and Bit-reversal traffic patterns are by about 38%, 44%, and 54% lower than by [22]. The network's saturation point of 2D-Mesh has been increased under Bit-reversal traffic. This is because the source nodes only send the packets to the destination nodes which are on the same subnet. Therefore, the multi-hops transmission between the long-distance nodes is almost low. Fig. 6 shows the saturation throughput (flits/node/cycle) of different architectures under synthetic workloads. The proposed method increase throughput in comparison to the VC based WiNoC and [22] under different traffic patterns.

The total energy consumption of the proposed method under random traffic for different packet injection rate is compared with VC based WiNoC and [22] in Fig. 7. This parameter is the sum of static and dynamic energy. As can be observed, the proposed network has a significantly lower energy consumption compared to VC based WiNoC, and also improved from [22]. This is due to the fact that for high loads, more flits can be transferred by the single-hop bypass channel and it leads to reduce energy consumption. Fig. 8 shows the total energy consumption under different traffic scenarios. It can be seen that the proposed scheme takes up smaller energy consumption among the three WiNoCs.

To evaluate the cost of the proposed microarchitecture, we can compare the area and power of the proposed baseline router and wireless router with other architectures. The proposed routers implement in RTL and synthesized using Synopsys Design Compiler in TSMC 45nm technology. The operating frequency is set to be 2 GHz. It is true that in both proposed routers, additional logic is needed to support

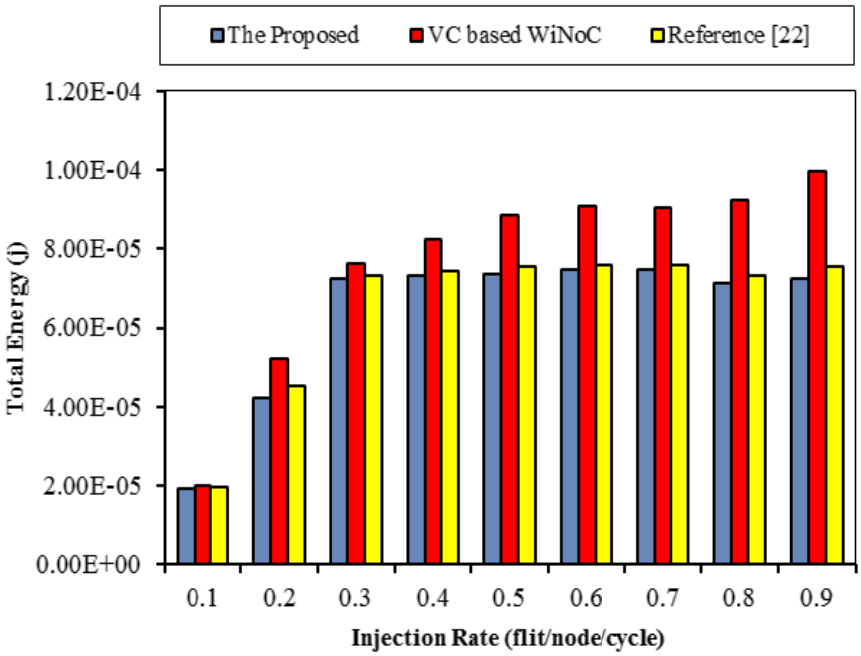

Fig. 7. Total energy consumption under Random traffic patterns for different packet injection rates.

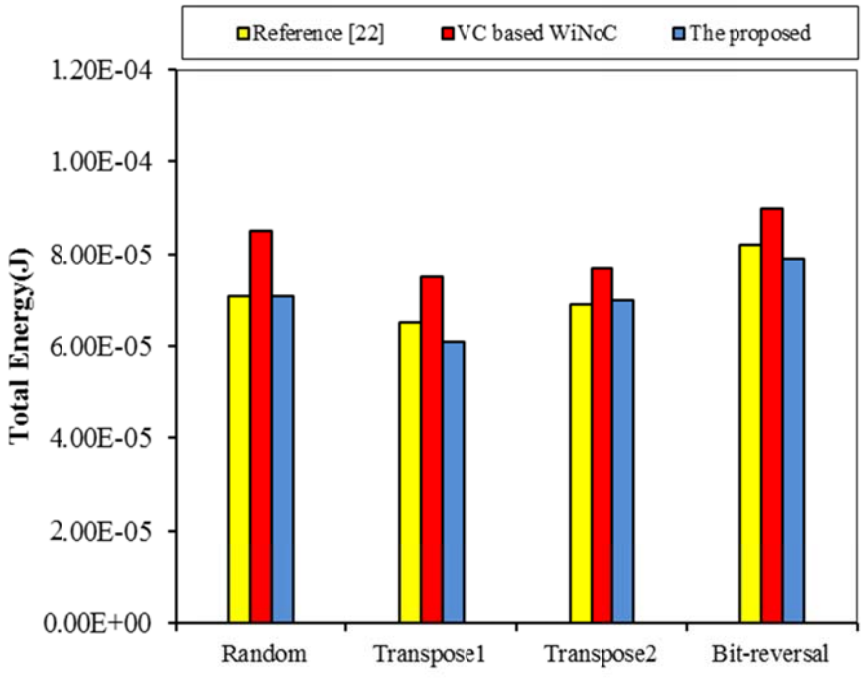

Fig. 8. Total energy consumption under different traffic scenarios.

Table VI. Routers area overhead.

WiNoC Architectures	**BR (um^2)** Total	**WR(um^2)**			**Critical path delay (ns)**
		Buffer	WI and others	Total	
WiNoC_VC	**119,899**	**137,756**	**76,970**	**214,726**	**0.23**
Reference [22]	**119,089**	**126,392**	**76,475**	**202,867**	**0.154**
The proposed	**120,494**	**126,392**	**79,075**	**205,467**	**0.063**

the bypass channels, but the area consumed by it is nearly negligible. However, some amount of power is consumed.

In spite of the reduction of input VCs in the proposed WRs, it can be seen that the area of the routers is higher than other WiNoC routers. In the proposed baseline router, the router's area increases by 1.18% compared with [22] mainly due to bypassing overhead. However, the critical paths have been removed in the proposed routers. The critical path in each router is often the control signals that drive the data path (e.g. mux select) and causes a large delay. This critical path is removed in the bypass routers because a pre-determined path is allocated to packets. Area overhead and critical path delay for all schemes are shown in Table VI.

IV. CONCLUSION AND FUTURE WORK

WiNoC is a novel interconnection architecture that can reduce the communication delay between the multi-hop long distant nodes. In WiNoC, wired and wireless routers transmit data packets to the short and long-distance routes. Obviously, when a large number of packets are sent to the network, congestion may occur and a number of routers are converted to the hot spot nodes. This problem decreases the utility of wired and wireless resources and can lead to the HOL blocking in the input ports of routers. In this study, the bypass channels and multiple VOQ were used in both routers to prevent the HOL blocking that can result in the reduction of network latency. Furthermore, an adaptive routing algorithm was proposed that use the advantages of low-latency bypass channels in the both routers. The simulation results show that the proposed scheme compared to other models can enhance the performance of the network with a significant reduction in the average latency of the network and an increase in the saturation throughput. As a part of the future work, it is suggested that the proposed scheme can be used to investigate the performance of real traffic pattern with a fault tolerance adaptive routing algorithm.

REFERENCES

[1] W. R. Davis *et al.*, "Demystifying 3D ICs: The pros and cons of going vertical," *IEEE Design & Test of Computers,* vol. 22, no. 6, pp. 498-510, Nov.-Dec. 2005.

[2] D. Vantrease *et al.*, "Corona: System implications of emerging nanophotonic technology," in *ACM SIGARCH Computer Architecture News*, vol. 36, no. 3, pp. 153-164, 2008.

[3] M. F. Chang *et al.*, "CMP network-on-chip overlaid with multi-band RF-interconnect," *14th International Symposium on High-Performance Computer Architecture,* pp. 191-202, 2008.

[4] A. B. Achballah, S. B. Othman, and S. B. Saoud, "Problems and challenges of emerging technology networks− on− chip: A review," *Microprocessors and Microsystems,* vol. 53, pp. 1-20, August 2017.

[5] B. A. Floyd, C.-M. Hung, and K.K. O "Intra-chip wireless interconnect for clock distribution implemented with integrated antennas, receivers, and transmitters," *IEEE Journal of Solid-State Circuits,* vol. 37, no. 5, pp. 543-552, May 2002.

[6] J.-J. Lin, H.-T. Wu, Y. Su, L. Gao, A. Sugavanam, and J. E. Brewer, "Communication using antennas fabricated in silicon integrated circuits," *IEEE Journal of solid-state circuits,* vol. 42, no. 8, pp. 1678-1687, August 2007.

[7] K. Kempa *et al.*, "Carbon nanotubes as optical antennae," *Advanced Materials,* vol. 19, no. 3, pp. 421-426, 2007.

[8] S.-B. Lee *et al.*, "A scalable micro wireless interconnect structure for CMPs," in *Proceedings of the 15th annual international conference on Mobile computing and networking*, pp. 217-228, 2009.

[9] Q.-T. Vien, M. O. Agyeman, T. A. Le, and T. Mak, "On the nanocommunications at THz band in graphene-enabled wireless network-on-chip," *Mathematical Problems in Engineering,* 2017.

[10] D. Zhao and Y. Wang, "SD-MAC: Design and synthesis of a hardware-efficient collision-free QoS-aware MAC protocol for wireless network-on-chip," *IEEE Transactions on Computers,* vol. 57, no. 9, pp. 1230-1245, Oct. 2008.

[11] D. DiTomaso, A. Kodi, D. Matolak, S. Kaya, S. Laha, and W. Rayess, "A-WiNoC: Adaptive wireless network-on-chip architecture for chip multiprocessors," *IEEE Transactions on Parallel and Distributed Systems,* vol. 26, no. 12, pp. 3289-3302, Dec. 2015.

[12] W.-H. Hu, C. Wang, and N. Bagherzadeh, "Design and analysis of a mesh-based wireless network-on-chip," The *Journal of Supercomputing,* vol. 71, no. 8, pp. 2830-2846, August 2015.

[13] A. Rezaei, M. Daneshtalab, and D. Zhao, "CAP-W: Congestion-aware platform for wireless-based network-on-chip in many-core era," *Microprocessors and Microsystems,* vol. 52, pp. 23-33, 2017.

[14] K. K. Chidella and A. Asaduzzaman, "A novel Wireless Network-on-Chip architecture with distributed directories for faster execution and minimal energy," *Computers & Electrical Engineering,* vol. 65, pp. 18-31, Jan. 2018.

[15] C. Qiuli, X. Wei, D. Fei, and H. Ming, "A reliable routing protocol against hotspots and burst for UASN-based fog systems," *Journal of Ambient Intelligence and Humanized Computing,* pp. 1-13, May 2018.

[16] S. Mikaeeli Mamaghani and M. A. Jabraeil Jamali, "A load-balanced congestion-aware routing algorithm based on time interval in wireless network-on-chip," *Journal of Ambient Intelligence and Humanized Computing, pp. 1-14,* 2018.

[17] A. Rezaei, M. Daneshtalab, M. Palesi, and D. Zhao, "Efficient congestion-aware scheme for wireless on-chip networks," in *24th Euromicro International Conference on Parallel, Distributed, and Network-Based Processing (PDP),* pp. 742-749, 2016.

[18] J. Murray, R. Kim, P. Wettin, P. P. Pande, and B. Shirazi, "Performance evaluation of congestion-aware routing with DVFS on a millimeter-wave small-world wireless NoC," *ACM Journal on Emerging Technologies in Computing Systems (JETC),* vol. 11, no. 2, p. 17, Nov. 2014.

[19] A. Rezaei, M. Daneshtalab, D. Zhao, F. Safaei, X. Wang, and M. Ebrahimi, "Dynamic application mapping algorithm for wireless network-on-chip," in *23rd Euromicro International Conference on Parallel, Distributed and Network-Based Processing (PDP),* pp. 421-424, 2015.

[20] Y. Ouyang, Z. Li, K. Xing, Z. Huang, H. Liang, and J. Li, "Design of Low-Power WiNoC with Congestion-Aware Wireless Node," *Journal of Circuits, Systems and Computers,* vol. 27, no. 9, p. 1850148, 2018.

[21] S. T. Nguyen and S. Oyanagi, "A low cost single-cycle router based on virtual output queuing for on-chip networks," in *13th Euromicro Conference on Digital System Design: Architectures, Methods and Tools (DSD),* pp. 60-67, 2010.

[22] Y. Ouyang, J. Yang, K. Xing, Z. Huang, and H. Liang, "An improved communication scheme for non-HOL-blocking wireless NoC," *Integration,* vol. 60, pp. 240-247, Jan. 2018.

[23] A. F. Noghondar and M. Reshadi, "A low-cost and latency bypass channel-based on-chip network," *The Journal of Supercomputing,* vol. 71, no. 10, pp. 3770-3786, Oct. 2015.

[24] T. Krishna, C.-H. O. Chen, W. C. Kwon, and L.-S. Peh, "Breaking the on-chip latency barrier using SMART," in *19th International Symposium on High Performance Computer Architecture (HPCA),* pp. 378-389, 2013.

[25] M. Opoku Agyeman, W. Zong, A. Yakovlev, K.-F. Tong, and T. Mak, "Extending the Performance of Hybrid NoCs beyond the Limitations of Network Heterogeneity," *Journal of Low Power Electronics and Applications,* vol. 7, no. 2, p. 8, 2017.

[26] V. Catania, A. Mineo, S. Monteleone, M. Palesi, and D. Patti, "Cycle-accurate network on chip simulation with noxim," *ACM Transactions on Modeling and Computer Simulation (TOMACS),* vol. 27, no. 1, p. 4, Nov. 2016.