Wireless Networks (2019) 25:3675–3687
https://doi.org/10.1007/s11276-019-01962-3

A fault-tolerant and congestion-aware architecture for wireless networks-on-chip

Seyed Hassan Mortazavi[1] · Reza Akbar[1] · Farshad Safaei[1] · Amin Rezaei[2]

Published online: 21 February 2019

Abstract
The combination of traditional wired links for regular transmissions and express wireless paths for long distance communications is a promising solution to prevent multi-hop network delays. In wireless network-on-chip technology, wireless-equipped routers are more error-prone than the conventional ones not only because of their implementation complexities but also due to their relatively high utilization. In this paper, a new topology is presented to enhance the network reliability, and then a novel routing algorithm is proposed to tolerate both intermittent and permanent faults on wireless hubs. In the proposed approach, once a wireless hub becomes faulty, the best alternative adjustment hub will be indicated and all the packets that have high average hop-count are routed through this alternative hub. In comparison with the state-of-the-art works, the proposed approach shows significant improvements in terms of robustness, congestion management, and resilience.

Keywords Network-on-chip · Hybrid wireless network-on-chip · Many-core system-on-chip · Reliability · Robustness · Congestion control management

1 Introduction

Multi-objective goals such as low-power, high-reliability, and low-temperature demands have forced the electronic industry to migrate from single-core systems to multi and many-core ones. Therefore, the communication efficiency of these systems has gain a lot of importance [1]. In addition, modern complex systems are being implemented on Networks-on-Chip (NoC) infrastructure in which the on chip communication is more critical than the traditional bus-based systems [2]. A lot of fault-tolerant and congestion-aware works such as [3, 4] have been done on two-dimensional NoC to improve performance. On the other hand, different new solutions with their advantages and disadvantages have been developed to improve the network performance and reduce the average hop-count. Such solutions include three-dimensional NoC [5–8], optical NoC [9, 10], NoCs based on RF communication channels [11], and wireless NoC [12–15].

Hybrid Wireless Network-on-Chip (HWNoC) has gained a lot of attention in recent years because of its unique features such as eliminating the long wire communication and providing irregular and reconfigurable topologies. The communication in HWNoC is done via a combination of wired and wireless links for short and long distance communications respectively. An HWNoC architecture consists of different parts such as wireless transceivers and their placement, network topologies, routing algorithms, shared wireless media access control, and task mapping [16]. Despite the advantages of HWNoC, there are challenges that we will tackle in this paper such as network reliability and wireless channel efficiency.

✉ Farshad Safaei
f_safaei@sbu.ac.ir

Seyed Hassan Mortazavi
info@hassanmortazavi.ir

Reza Akbar
r_akbar@sbu.ac.ir

Amin Rezaei
me@aminrezaei.com

[1] Faculty of Computer Science and Engineering, Shahid Beheshti University G.C., Evin, Tehran 1983963113, Iran

[2] Department of Electrical Engineering and Computer Science, Northwestern University (NU), Evanston, USA

Given the need for reliability in digital systems, an important issue in designing interconnections for large-scale multiprocessor architectures is fault tolerance that indicates the operational ability of the network even with defective components [5]. In different systems, faults are divided into three categories: transient, intermittent, and permanent. Transient faults occur accidentally for one or several clock cycles and temporarily affect the functionality of the system for a short time. In HWNoC, the source of these faults can be environmental noises in the wireless media. Intermittent faults occur during a certain time under an undesirable condition; the system does not function properly until the element returns to the favorable condition. Intermittent faults occur in HWNoC because of the frequent use of wireless hubs and links. On the other hand, permanent faults occur when a defective element can no longer function; in this case, the system must be adapted quickly to avoid performance degradation [7].

Wireless-equipped hubs have higher failure rates than the other chip components due to their implementation complexities as well as their relatively high utilization [17, 18]. This makes the need of a fault-tolerant platform in HWNoC more critical than their wired counterparts. The main goal of this paper is to provide a fault-resistant HWNoC architecture against both the intermittent and the permanent faults. On the other hand, an efficient way to improve the HWNoC performance is to implement routing algorithms that consider wireless channel utilities and balance the usage of the wired and the wireless links. Thus, we introduce a novel topology that considers the wireless hubs failure combined with cost efficiency of the network. Then, we propose a novel congestion-aware routing algorithm to improve the network performance.

The remainder of the paper is organized as follows. In Sect. 2, we introduce existing HWNoC architectures. In Sect. 3, an improved topology with novel fault-tolerant architecture is proposed. The experimental results are shown in Sect. 4. Finally, Sect. 5 delivers conclusions and suggestions for future works.

2 Related work

In HWNoC, the combination of the wired and the wireless links can be utilized to implement different topologies. Among them, small world networks with multi-dimensional and hierarchical architecture reduce the network diameter and ultimately improve the performance [12–15]. It is proved that a small world network with n nodes will have a diameter corresponding to $log\ n$ [14]. In these networks, shortcuts are used to communicate between different subnets; in this case, although they have fewer resources compared to the complete networks, they have higher efficiency in comparison with the conventional mesh networks [19]. However, the high degree of connection in the hubs and communication bridges located between the different hierarchical levels of the graph causes excessive overhead in the small world topology. In addition they may create hot-spot due to their shared nature. When the hubs fails, the subnet attached to that hub will be completely disconnected from the network. Another topology that is used in HWNoC, considers a global mesh network as baseline and equips some of the nodes to wireless transceivers [20–25]; in this case, the wireless transceiver placement plays an important role in the network performance. Based on the obtained results from the above works, networks with small world topologies have better performance while there is no network fault. However, they spend more area and power to buy performance.

Ultra Wide Band (UWB) based transceivers that are used in [26] require multi-hop packet transfer due to their short transmission bandwidth, which reduces the overall network performance. An HWNoC based on the terahertz waves is described in [12]. Although it has 24 channels of 10-gigabit-per-second between wireless transceivers, it faces integration problems due to the need for a direct line of sight between these transceivers. Thus, millimeter wave (mm-Wave) transceivers with zigzag antennas are proposed [13, 14]. These transceivers provide higher performance because of utilizing 16-gigabit-per-second channel with bit error rates of less than 10^{-15} and 20 mm range data transfer capability. In addition, mm-Wave technology does not need a direct line of sight to operate properly. An informative comparison between different wireless on-chip transceivers is shown in [27].

Given the close relationship between the routing algorithms of a network and its topology, different HWNoC-based routing algorithms are proposed. In hierarchical and small world network [12, 14], routing is performed in agreement with its subnet topology. If the destination of a packet is outside of the current subnet, it will first be directed to the subnet hub and then forwarded to the destination subnet hub, and finally to the destination node. Since only some of the hubs have a wireless link, routing in the second phase should be chosen smartly. Moreover a Token Flow Control (TFC) mechanism is designed to avoid the creation of hotspots in the hubs equipped with the wireless interfaces [12]. The token is obtained based on the number of flits enters the input port buffer of the hub with the wireless interface.

On the other hand, in mesh topologies, the entire network is divided into subnets; hence, routing is done according to the subnet location of the source node and the destination node. Upon arriving the packet's header to the router, if the destination of the packet is in the current subnet, routing is performed through the conventional

routers without involving the wireless channel. Otherwise, paths between the source and the destination with and without using the wireless links are calculated and the shortest one is chosen when the destination node is in another subnet. Since the shortest path is usually the path containing the wireless routers, they can rapidly become hotspots. In order to solve this problem, the routing algorithms proposed in [20–25] implement a parameter δ that determines routing from a wired or wireless route. The larger δ means only the packets with a long distance between source and destination use the wireless channel. In [21], this parameter has a fixed value of 6, while in [22–25] this parameter is dynamically measured and calculated according to the network utilization rate of the wireless channel as well as the ratio of wireless routers to the all the routers. However, these routing algorithms did not consider the fault tolerance mechanisms.

One of the important components of modern NoCs is the Media Access Control (MAC) mechanism that can be FDMA, TDMA, CDMA or competition-based mechanisms and CSMA/CA. Among them, FDMA and TDMA have less complexity and area overhead and thus more popular in HWNoC. In [12] simply by partitioning the channels between the on-chip wireless links and the FDMA mechanism, the system requirements for non-disturbance in media are met. In [13, 14], the single-channel constraint has created a need for an advanced MAC mechanism. Thus, a TDMA-based token passing mechanism is used in which the hubs equipped with wireless transceiver are connected via the ring topology. In the token passing mechanism, only one transceiver can take control of a radio channel and sends information on it; all the other transceivers are set to the receiver mode. In this case, if the transmitter has no information to send, the token will be passed to the next hub to increase the productivity of the media [28]. In [29] the maximum number of clock cycles for the token to be possessed by a hub is determined dynamically based on the previous timelines. With the redistribution of unused clock cycles, the efficiency of the MAC mechanism is improved since the congested networks always have a higher chance to possess the token. In this paper, the later MAC mechanism is considered.

Moreover, Error Control Coding (ECC) in wireless communications is utilized in [30]. By evaluating the system in terms of various transient faults on a wireless channel, the resistance of HWNoC has been shown to be better than traditional NoCs. In addition, the performance of wireless network is improved in the presence of permanent faults in [17, 18] by utilizing auxiliary wireless nodes. In this paper, by moving the location of information packets to remote destinations, the migration from subnets with faulty wireless hubs and high-congestion is done to the subnets with non-faulty hubs and no congestion.

3 The proposed HWNoC architecture

The HWNoC architecture is characterized by the technology of the radio transceiver, the network topology, the routing algorithm, and the MAC mechanism. In this section, we are proposing a fault-tolerant HWNoC architecture along with its topology and routing algorithm. This architecture utilizes mm-Wave transceivers with zigzag antennas, adopts ECC mechanism for wireless hubs, and extends token ring wireless MAC mechanism.

3.1 Topology

To reconcile the topologies used in previous works, which are usually either mesh or small world, a combined topology has been used. In this topology, all the nodes are connected together in a large mesh, and the entire mesh is divided into several subnets. Instead of equipping some routers with wireless transceivers, separate hubs have been used in the subnets. This will increase the network's tolerance to faults compared to small world networks, and improve network performance in terms of delay and energy dissipation compared to wireless mesh-like networks equipped with wireless routers. Figure 1(a) shows a 256-core network. In this topology, the network is divided into 4×4 subnets. As can be seen, each subnet has a wireless hub. Figure 1(b) depicts a subnet of a wireless hub and 16 typical routers from the top view.

Similar to the small-world network, the proposed topology has small diameter. In hierarchical architectures, the wireless hubs only broadcast the network flits on wireless channels while in this paper, the wireless hubs generate packets. In addition the topology of all the subnets is a wide mesh. In the case of a hub failure, the packets within the subnet of that hub can be routed in the first level wired mesh.

Now, we are going to introduce a platform for producing an efficient HWNoC topology in which the reliability of the wireless hubs is considered as an input parameter. In each subnet, the wireless hub can be connected to any subset of 16 different routers. As the number of routers connected to the hub is increased, the number of wireless hub's ports will be increased. Increasing the number of ports not only increases the occupied area but also causes congestion due to the high-utilization of the wireless network. Therefore, after a certain point, the increase in wireless utilization will be useless. On the other hand, increasing the number of wireless hubs reduces the diameter of the network. Thus, there is a trade-off between the number of wireless hubs and the system performance. For finding an efficient trade-off, Simulated Annealing (SA) algorithm will be used.

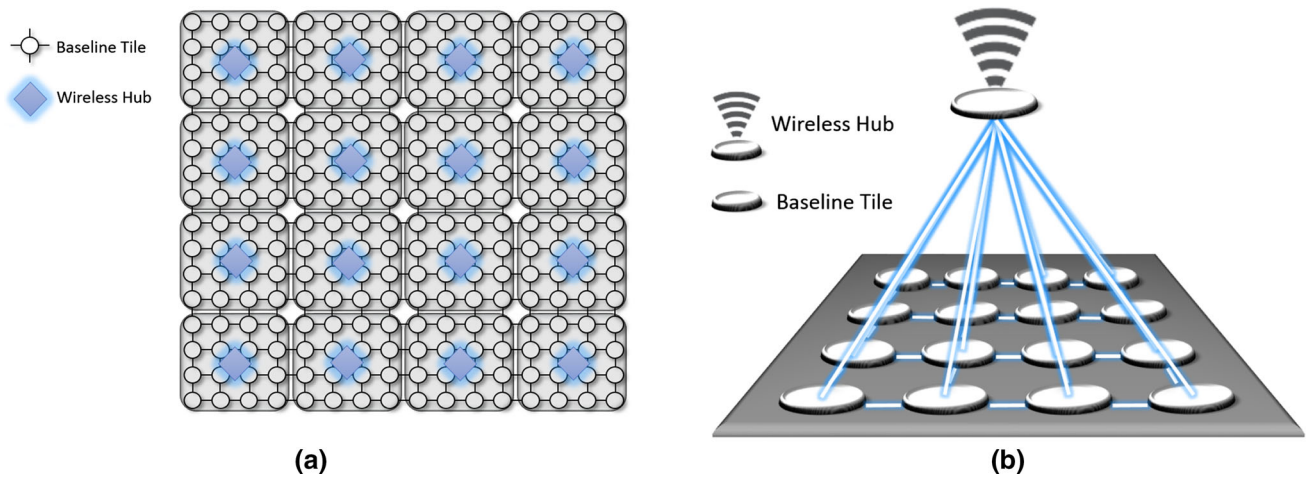


Fig. 1 **a** A 256-node network with 16 subnet each equipped with a wireless hub, **b** a mesh subnet with 16 conventional routers and one wireless hub

SA is a simple meta-heuristic algorithm for solving optimization problems in large search spaces. By making minor changes in the state space of the main problem, SA attempts to improve the cost function and thus reach a near-optimal point. The cost function considered for SA is given by

$$\text{cost} = \alpha \prec \tilde{d} \succ + (1 - \alpha) \cdot \tilde{\Lambda} \tag{1}$$

in which $\prec \tilde{d} \succ$ specifies the ratio of the average distance between all pairs of nodes in the network with wireless hubs to the average distance between all pairs of nodes in a mesh without any wireless hub. This parameter is normalized between zero and one. Further, the variable $\tilde{\Lambda}$ denotes the ratio of area overhead in the network with wireless hubs to the maximum possible area overhead. The maximum area overhead occurs when a wireless hub is considered for each node in the network. This parameter is also normalized between zero and one. And finally, the parameter α is a number between zero and one. Once it is closer to one, the importance of the mean distance increases and once it is closer to zero, overhead becomes more important.

As stated above, to calculate $\prec \tilde{d} \succ$, we need to calculate the average distance between all pairs of nodes in the new network and the fully wired conventional mesh network. To obtain the mean distance between nodes, the number of hops between every pair of nodes should be averaged. That is

$$\prec \bar{d} \succ = \frac{\sum_{i,j} (d_{\prec i,j \succ} \cdot T_{\prec i,j \succ})}{\sum_{i,j} T_{\prec i,j \succ}} \tag{2}$$

where $d_{\prec i,j \succ}$ is the shortest distance between the two nodes i and j, and $T_{i,j}$ represents their traffic volume between the two nodes.

Obtaining an average distance or number of hops in a network is a simple action if the presence or absence of all links is definite. However, we must get this average distance considering the possibility of a wireless hub failure in an HWNoC. If we assume that, the coordinates of the nodes in the topology of M × N mesh start from the left bottom, in order to obtain the shortest distance between the two nodes i and j, the following equation can be written when there is no wireless link in the network. The shortest distance between i and j, the two nodes in the conventional M × N mesh network without any wireless links is obtained through Eq. (3). Note that the nodes in the network are numbered from 0 to $M \times N - 1$.

$$d_{\prec i,j \succ, \bar{h}} = |i \bmod M - j \bmod N| + ||\lfloor i/M \rfloor - \lfloor j/N \rfloor|| \tag{3}$$

Additionally, the routing is possible through wireless hubs between different network nodes. As a result, the nodes distance can be reduced to these hubs, which are quick links and shortcuts in the network. To calculate the distance between nodes i and j using the intermediate wireless hubs, we get

$$d^{h}_{d_{\prec i,j \succ}} = d_{\prec i,S \succ} + d_{\prec S,D \succ} + d_{\prec D,j \succ} \tag{4}$$

where $d^{h}_{d_{\prec i,j \succ}}$ is the shortest distance between the two nodes i and j utilizing wireless hubs, if both of the source and the destination wireless hubs are present and healthy. Also, $d_{\prec i,S \succ}$ and $d_{\prec D,j \succ}$ respectively, show the distance between i and the wireless hub in the source subnet and the distance between wireless hubs in the destination subnet to j. Finally, the parameter $d_{\prec S,D \succ}$ specifies the distance between wireless hubs in the source and the destination subnets. If the node is directly connected to the wireless hub, the distance between node and wireless hub will be equal to one; otherwise, according to Eq. (3), the packet routes to the nearest connected node to the wireless hub

using the wired mesh network, and then enters to the wireless hub from there. The distance calculation between two wireless hubs will be explained in the following of the manuscript.

In the case of a presence/absence of failure in one or both wireless hubs in the source and the destination subnets, nodes i and j can use adjacent subnet hubs. In the following, we plan to calculate the distance between the two wireless hubs. Despite the fact that the number of healthy wireless hubs in the network is known, the distance between the two hubs is inversely proportional to their number. The distance between the two wireless hubs in an HWNoC is calculated by

$$d_{\prec h,h\succ} = \frac{FS \cdot \sum_{i=1}^{N_h} n_i}{WD \cdot CP \cdot N_h} \tag{5}$$

where $d_{\prec h,h\succ}$ is the distance between two wireless hubs in the network and FS denotes the size of a flit in terms of bits. Moreover, N_h indicates the number of wireless hubs in the network and n_i is the number of wired input ports for hub i. WD and CP are the bandwidth of the wireless channels and the clock period respectively.

Accordingly, the quantity $\frac{FS}{WD \cdot CP}$ indicates the number of hubs that a flit needs to transfer from one hub to another via a wireless channel, assuming that the wireless channel is completely empty and free. In Eq. (5), the distance between the two wireless hubs has a relation with the bandwidth of the wireless channel used in the network. Therefore, using an appropriate MAC mechanism can ultimately optimize the distance between the two hubs. In addition, to decide on the use of wireless channels when routing different packets, calculating and measuring the distance between two wireless hubs is important.

The shortest distance between i and j using wireless hubs is calculated according to Eq. (6) with considering the reliability of wireless hubs at the time of designing and building HWNoC.

$$\begin{aligned} d_{\prec i,j\succ,h} = {} & R_S R_D \cdot d^{SD}_{\prec i,j\succ} + \bar{R}_S R_D \cdot d^{\bar{S}D}_{\prec i,j\succ}\{1-(\bar{R}_S)^x\} + R_S \bar{R}_D \\ & \cdot d^{S\bar{D}}_{\prec i,j\succ}\{1-(\bar{R}_D)^y\} + \bar{R}_S \bar{R}_D \\ & \cdot d^{\bar{S}\bar{D}}_{\prec i,j\succ}\{1-(\bar{R}_S)^x\}\{1-(\bar{R}_D)^y\} + d_{\prec i,j\succ} \cdot P_w \end{aligned} \tag{6}$$

in which $d_{\prec i,j\succ,h}$ is the distance between the two nodes i and j using a faulty or healthy wireless hubs. In addition R_S and R_D show the reliability of the source and the destination wireless hubs respectively. We assumed that these two values are equal and follow the exponential distribution. So, they are calculated by $R_S = R_D = e^{-\lambda t}$. Thus, $\bar{R}_S$ and $\bar{R}_D$ are considered as the probability of having fault in the source and the destination wireless hubs. The parameter $d^{\bar{S}D}_{\prec i,j\succ}$ in Eq. (6) denotes the shortest distance between nodes i and j, assuming that the source hub is faulty and that the destination hub is healthy. Similarly, other parameters can be interpreted based on the failure of one or both source and destination hubs. Additionally, x and y indicate the number of hubs adjacent to the faulty source and destination hubs respectively. Finally, the P_w quantity implies the probability of using a wired path. The probability of this quantity is equal to the probability of failure in the source hub and all adjacent hubs, or the probability of failure in the destination and all adjacent hubs. Therefore, we have

$$P_w = \Pr\{E_1 \cup E_2\} \tag{7}$$

where the events E_1 and E_2 are defined as

$$E_1 = (\bar{R}_S)^{x+1}, E_2 = (\bar{R}_D)^{y+1} \tag{8}$$

Increasing the reliability of wireless hubs in Eq. (6) will increase the probability of using them in the routing algorithm, which reduces the distance between two nodes. Obviously, as the number of hubs adjacent to the hub in the source or destination subnet increases, the fault tolerance of the entire network should increase as well.

Finally, to find the distance between each pair of nodes, we must take the minimum amount of the Eq. (6) and (3). Note that the use of wireless hub may increase the distance under the circumstances. Thus, we have

$$d_{\prec i,j\succ} = \mathrm{Min}\{d_{\prec i,j\succ,\bar{h}}, d_{\prec i,j\succ,h}\} \tag{9}$$

where $d_{\prec i,j\succ,\bar{h}}$ and $d_{\prec i,j\succ,h}$ show the distance between the two nodes i and j in network with or without wireless hubs.

According to Fig. 2, if the source or destination of a packet is in a subnet, which is located at the center of the failed hub's area then the second level network and wireless media cannot be used in routing.

In order to reduce the area of implementation of network hubs and increase the efficiency of the ports, the connection of some links between subnet nodes to the corresponding hub can be ignored as shown in Fig. 3. The number of hops required from each node in the subnet will increase to its subnet hub. In this figure, the distance between each node in the subnet and the wireless hub is two.

Now we want to explain how to calculate $\tilde{\Lambda}$ in Eq. (1). To obtain the area overhead of a wireless hub we have

$$\Lambda_i = n_i \cdot A_i + A_r, \quad i = 1, 2, \ldots, N_h \tag{10}$$

where Λ_i is the area overhead of wireless hub i and n_i is the number of wired input ports for that hub. Also, A_i is the area needed for buffers and all components of a port in a wireless hub and A_r denotes the area needed to implement the radio transceiver interface used in the wireless hub, which is specified before the implementation. The

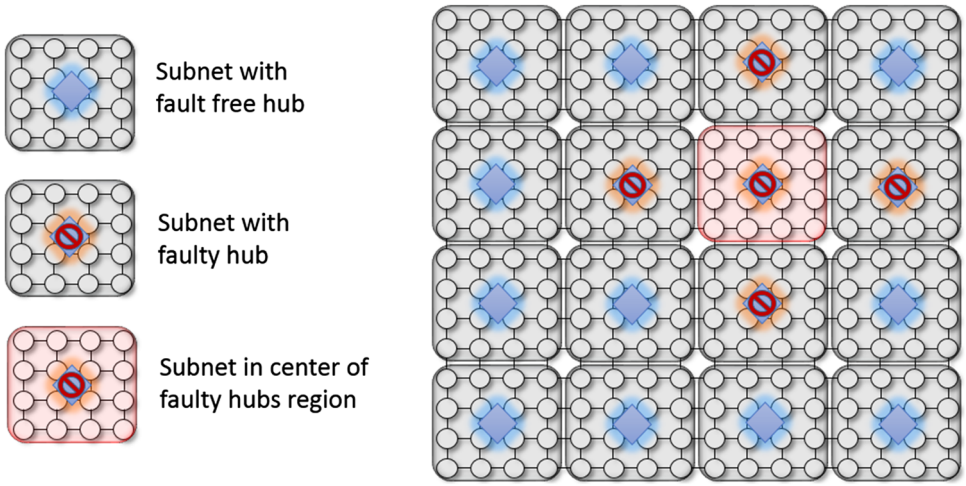


Fig. 2 A wireless network-on-chip, which has a subnet in the center of the area with five failed wireless hubs

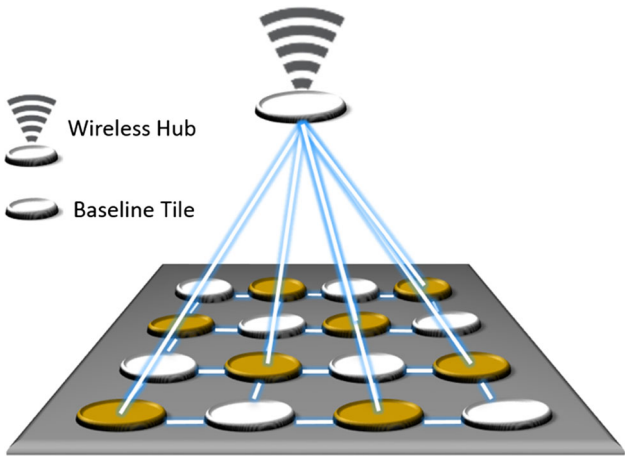


Fig. 3 An example of how to connect a wireless hub to the nodes of a subnet provided

overhead area of wireless hubs in the entire HWNoC, Λ_t, can be written as

$$\Lambda_t = \sum_{i=1}^{N_h} \Lambda_i \tag{11}$$

In order to get $\tilde{\Lambda}$ in Eq. (1), we need to calculate the maximum area overhead of wireless hubs in the NoC. This happens when assuming each node of graph as a subnet, and for each node, we consider a wireless hub. Assuming having an NoC with mesh topology of the size M × N, the maximum area overhead,Λ_m, is obtained by

$$\Lambda_m = \sum_{l=1}^{M} \sum_{k=1}^{N} \underset{i=1}{\overset{N_h}{\text{Max}}} \{A_i + A_r\} = M.N. \underset{i=1}{\overset{N_h}{\text{Max}}} \{A_i + A_r\} \tag{12}$$

To capture the maximum area overhead, we must get the maximum overhead per hub and then calculate it for all nodes in the network. Since there is a wireless hub for each node in calculating the maximum overhead, then the parameter n_i, which is the number of wired input ports for hub i, must be one

According to Eqs. (11) and (12), it is easy to calculate the hub area overhead ratio,$\tilde{\Lambda}$ required by Eq. (1). Then, we get

$$\tilde{\Lambda} = \Lambda_t / \Lambda_m \tag{13}$$

The lower the value of the calculated ratio in Eq. (13) is, the better the graph obtained with respect to the area. Clearly, the best case is when the topology is a wide mesh without any wireless hub (i.e., α equals zero in Eq. (1). Table 1 lists a summary of the symbols used in the model, along with their interpretation.

3.2 Routing algorithm

To achieve the best performance of the improved HWNoC architecture presented in the previous sub-section, as well as improving the resilience of the wireless hubs, we intend to introduce an adaptive fault-tolerant wireless routing algorithm. The main idea of the proposed algorithm is based on the fact that by reporting the occurrence of a fault or congestion in the wireless hub of each subnet, the packets in that subnet will be directed to the closest subnet with healthy wireless hub. The fault report can be generated through a Built-In Self-Test (BIST) system embedded and distributed on wireless hubs.

Figure 4 shows a network with 256 nodes connected to each other as a 16 × 16 mesh. Each 16 nodes form a subnet and all the nodes in each subnet communicate with a wireless hub in that subnet. According to the figure, if node 4 of subnet 4 intends to send a packet to node 7 of subnet 10, it can reach via different routes to its destination. Assuming a wireless and flawless interface in subnet hubs 4 and 10, the shortest route along the use of wireless channel

Table 1 List of symbols used in the analytical model

Symbol	Description
$\prec \tilde{d} \succ$	The ratio of the average distance between all pairs of nodes in the network with wireless hubs to the average distance between all pairs of nodes in a mesh without any wireless hub
$\tilde{\Lambda}$	The ratio of area overhead in the network with wireless hubs to the maximum possible area overhead
α	A number between zero and one; once it is closer to one, the importance of the mean distance increases and once it is closer to zero, the overhead becomes more important
$\prec \bar{d} \succ$	Mean distance between all pairs of nodes in the network
$d_{\prec i,j \succ}$	The shortest distance between two nodes i and j
$T_{\prec i,j \succ}$	The traffic volume between two nodes i and j
N	The height of the mesh topology
M	The width of the mesh topology
n_i	The number of the wired input ports for hub i
P_w	The probability of utilizing a wired path
$d^h_{d_{\prec i,j \succ}}$	The shortest distance between two nodes i and j considering wireless hubs, if both source and destination wireless hubs are present and healthy
$d_{\prec i,S \succ}$	The distance between i and the wireless hub in the source subnet
$d_{\prec S,D \succ}$	The distance between the wireless hubs in the source and the destination subnets
$d_{\prec D,j \succ}$	The distance between the wireless hub in the destination subnet and node j
$d_{\prec h,h \succ}$	The distance between two wireless hubs in the network
FS	The size of a flit in terms of bits
N_h	The number of wireless hubs in the network
WD	Bandwidth of the wireless channels
CP	Clock period
$d_{\prec i,j \succ,h}$	The distance between two nodes i and j using a faulty/healthy wireless hub
$d_{\prec i,j \succ,\bar{h}}$	The shortest distance between two nodes i and j in the conventional mesh network without any wireless links
R_S	The reliability of the source wireless hub
R_D	The reliability of the destination wireless hub
$\bar{R}_S = 1 - R_S$	The failure probability in the source wireless hub
$\bar{R}_D = 1 - R_D$	The failure probability in the destination wireless hub
$d^{SD}_{\prec i,j \succ}$	The shortest distance between two nodes i and j, assuming that the source hub is faulty and that the destination hub is healthy
Λ_i	The overhead area of a wireless hub
A_i	The area needed for buffers and all components of a port in a wireless hub
A_r	The area needed to implement the radio transceiver interface used in the wireless hub
Λ_t	The area overhead of wireless hubs in entire HWNoC
Λ_m	The maximum area overhead
Φ_{cur}	The address of the current node
M_s	The width of the subnet
N_s	The height of the subnet
Θ_i	A normalized parameter between zero and one. Index i that can be R, L, U, or D is indicated to the right, left, up, or down, respectively
h	Total number of subnets in the first dimension (i.e., vertical)
v	Total number of subnets in the second dimension (i.e., horizontal)
Ψ_{sub}	Subnet address that does not have a wireless hub or have a faulty hub
Ω_i	A Boolean variable. The index i can be R, L, U, or D that is indicated to the right, left, up, or down, respectively

is routing. Therefore if you consider the number of hops required to send on a wireless channel one hop, each flit of this packet needs 3-hubs to arrive at the destination. Now, with the assumption of the wireless hub failure in subnet 4, the shortest route to deliver the packet on the destination is to use the wireless hub of the closest subnet. As can be seen

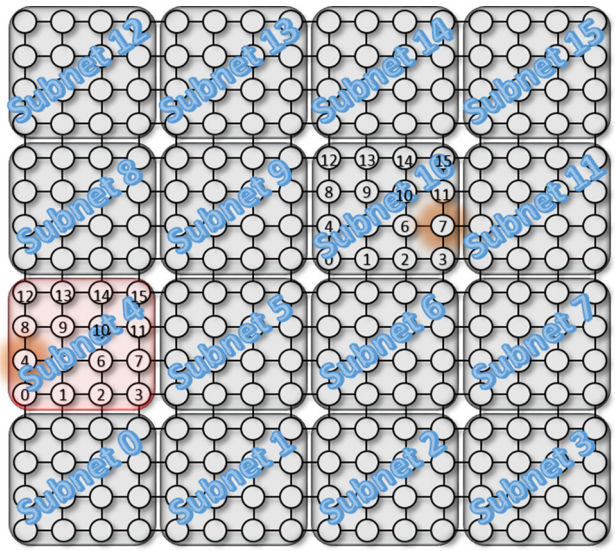


Fig. 4 16 × 16 HWNoC architecture

in the figure, the closest subnet to node 4 of subnet 4 is the subnet zero. In this case, each flits requires five hops from the source to the destination. It should be noted that if we use a typical mesh network with no wireless hubs, the minimum number of hops needed for each flit, will be increased to 15.

As shown above, the existence of a suitable routing algorithm causes a large difference in the number of required hops between the source and the destination. If each packet were composed of multi-flits with assuming a probability of having these connections over the entire network operation time, the difference in the choice of these two routes and its effect on the performance of the entire system would be realized. Thus, we propose a distributed routing algorithm that considers the situation in which the wireless hubs become faulty or lacking wireless hub in the subnet.

Assuming that the target subnet has M_s nodes in the first dimension and N_s nodes in the second dimension, Eq. (14) represents the decision weight of the router output ports to determine the route priority of the flit header. According to the wormhole switching, the rest of the flits follow the header flit until the entire packet reaches its destination. In Eq. (14), the parameter Φ_{cur} denotes the address of the current node where the routing decision is made, and is located in a subnet without wireless hub or with faulty wireless hub.

$$\begin{cases} \Theta_R = \dfrac{\Phi_{cur} \bmod M_s}{M_s - 1} \\ \Theta_L = 1 - \Theta_R \\ \Theta_U = \dfrac{\lfloor \Phi_{cur}/N_s \rfloor}{N_s - 1} \\ \Theta_D = 1 - \Theta_U \end{cases} \tag{14}$$

Also, Θ_i is a normalized parameter that has a value in the range of [0, 1]. The indices of R, L, U, and D are indicated right, left, up and down, respectively. The relationships between calculating decision weight for the edges of the wide mesh should be different from the aforementioned relations. In this case, in the current router only the weight of the ports is calculated, which has the necessary condition for the port on that router in accordance with the direction of the port. These conditions are given in Eq. (15).

$$\begin{cases} \Omega_R \leftarrow (\Psi_{sub} \bmod h \neq h - 1) \\ \Omega_L \leftarrow (\Psi_{sub} \bmod h \neq 0) \\ \Omega_U \leftarrow (\Psi_{sub} < h \cdot v - h) \\ \Omega_D \leftarrow (\Psi_{sub} \geq h) \end{cases} \tag{15}$$

In the above relationships, Ψ_{sub} is the subnet address of damaged or lacking a wireless hub over the entire network. Parameters h and v also represent the total number of subnets in the first dimension and the total number of subnets in the second dimension in the entire network. In Fig. 4, Ψ_{sub} value of a damaged subnet or lacking a wireless hub is 4, and both h and v are equal to 4 in the entire network. Ω_i is also a Boolean variable.

Figure 5 illustrates a partial network of Fig. 4 to deal with the mentioned scenario in the network. Considering the node 4 of subnet 4, the routing algorithm should choose to navigate via the south port when encounters a fault in the subnet 4 wireless hub because, the goal is to route the packet to the closest subnet, which is the subnet zero here.

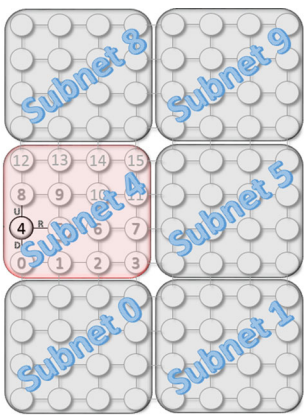


Fig. 5 An example of the proposed algorithm

By inserting the index number of the source subnet (i.e., 4), as well as the index number of the source node (i.e., 4) in Eq. (15), we will get the expressions in Eq. (16). These relationships indicate that, when a network encountering with faults or congestion, the network transfers packets within this router to the south port toward the subnet zero.

$$\begin{cases} \Theta_R = \dfrac{4 \bmod 4}{4-1} = 0, \Theta_L = 1 - 0 = 1, \\ \Theta_U = \dfrac{\lfloor 4/4 \rfloor}{4-1} \simeq 0.33, \Theta_D = 1 - 0.33 = 0.66 \\ \Omega_R = \Omega_U = \Omega_D = 1, \Omega_L = 0 \\ \Theta_D > \Theta_U > \Theta_R \end{cases} \tag{16}$$

All of the above mechanisms are applicable during a fault or congestion on the hubs connected to source and destination subnets of packet and only for packets, which are distant from each other. When this happens, the routing provided in this paper is used; otherwise the routing is done according to previous different routing algorithms. For this reason, the routing proposed in this paper completes previous routings. It's quite clear that when the hubs, which are the wireless channel input in the network fail, the efficiency of the wireless channel is greatly reduced, resulting in reduced overall system performance. However, in previous work such as [12–15, 20–25], assuming the integrity of all hubs or the system components that are responsible for delivering packets to a wireless channel, certain solutions have been presented to fixate the wireless channel productivity. A very important point in all routing algorithms is the decision to use/not use the wireless channel in routing packets in the network. Due to the shared nature of wireless channels in an HWNoC, the amount of wireless channel efficiency has a significant impact on the overall performance of the network. In this paper, Eq. (5) is used to find the average number of hops needed to transfer between two wireless hubs.

4 Performance evaluation

In this section, we are going to look at the results of the research. Initially, with the help of a tool to improve the topology written in the Python programming language, we have achieved an improved topology over the previous architectures of the HWNoC according to the presented method in Sect. 3.1. Eventually, we implement this topology and the routing algorithm proposed in Sect. 3.2 in the extended version of the Noxim Clock cycle simulator [31]. Then, we have compared our proposed architecture with state-of-the-art works in terms of robustness, congestion management, and resilience. In all stages of the simulation, the mechanism is derived from [29] with some improvement in its access to wireless media.

Figure 6 shows the average distance between the network nodes in the topologies presented in [22, 24], and the improved topology in this paper in different configurations. In Fig. 6(a), the improved topology is compared with the proposed topology in [22]. This compares with 10×10 networks with 4 wireless hubs, 15×15 networks with 9 wireless hubs and 20×20 networks with 16 wireless hubs. The improved topology is also considered with the same number of nodes and the number of wireless hubs. As can be seen, the improved topology has less average distance in comparison with the topology in [22]. Generally, the improved topology for different dimensions and reliability capabilities improved the distance between the nodes of the network by 12% compared with [22]. In Fig. 6b, the improved topology is compared with the presented topology in [24]. The results

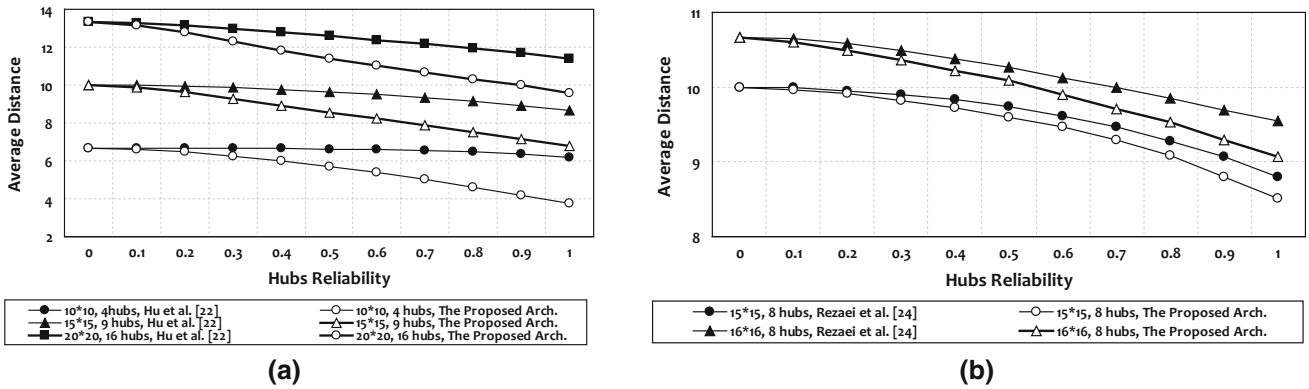


Fig. 6 Comparison of the average distance between different network nodes based on different injection rates in improved topology and **a** [22], **b** [24]

show that the improved topology reduces 2% of the average distance between the different nodes of the network compared with [24]. In all irregularly optimized topologies, using SA algorithm and enhancement tool, the reliability of wireless hubs is considered 90%. It should be noted that the reliability parameter considers both the wireless hubs and wireless links; thus, if we consider the reliability to be zero, the network transforms into a simple mesh.

Table 2 Simulation parameters

Parameter	Value
Number of nodes	64, 100, 256
Number of wireless hubs	4, 5, 8
Capacity of buffers connected to the wireless hub	4 flits
Capacity of router buffer	4 flits
Flit size	16 bits
System clock rate	1 GHz
Wireless channel characteristics	A 16 gigabits per second mmWave Band Transceiver with zigzag antenna [14] equivalent to 1 flits per cycle
Switching mechanism	Wormhole
Routing algorithms	Hue et al. [22], Rezaei et al. [24], Current work
Media access control mechanism	Extended token ring [29]
Wireless hubs reliability	0.8, 0.9, 1.0

Next, we consider the latency comparison. Table 2 shows the summary of the parameters used in the simulation configuration.

Figure 7 depicts the results of comparing the average delay in different HWNoC architectures given in Table 3 based on different injection rates and with different wireless hub reliabilities. These architectures consist of a routing and a topology. From the topology [22, 24], given that the percentage improvement in the proposed topology in this article is much different from [22], and [24] has shown a better performance. Thus, we only consider this topology in comparison of different network architectures. From the routing provided in [22, 24], we consider Wang routing as a routing for comparison. These two routing are very similar, with the difference that in [24]. The variable δ is dynamically calculated, and since the variable is dependent on the wireless link efficiency, which cannot be calculated at runtime due to the need of knowledge of the

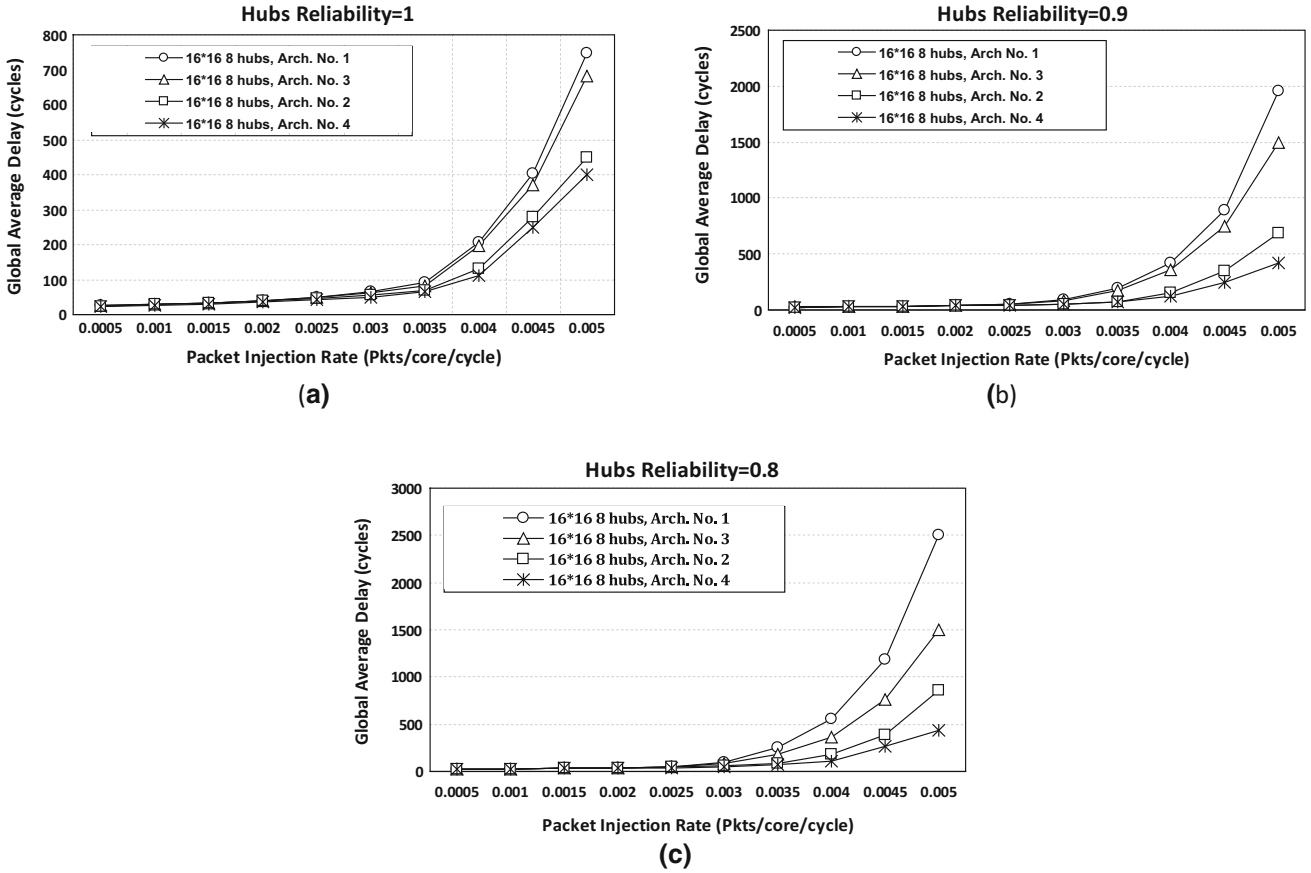


Fig. 7 Comparison of different HWNoC architectures at different injection rates

Table 3 Different designs for comparison

Architecture No.	Routing algorithm	Topology
1	Rezaei et al. [24]	Hu et al. [22]
2	Current work	Hu et al. [22]
3	Rezaei et al. [24]	Current work
4	Current work	Current work

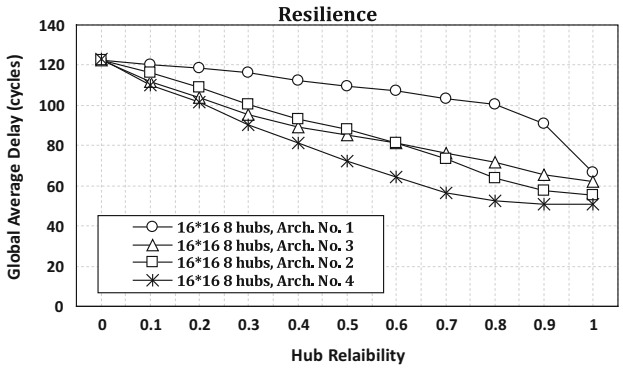


Fig. 8 Average delay in four different architectures for various reliability capabilities at the injection rate of 0.003

whole network, this dynamism is inefficient. Therefore, we consider topology [24] and routing [22] for comparison with the routing and topology presented in this article. Accordingly, we have two topologies and two routings, which together make up four network architectures. In summary, these four architectures are shown in Table 3.

In Fig. 7(a), the wireless hub reliability is assumed to be one. In this case, all wireless hubs will be functional. The topology under consideration is 16 × 16 nodes with eight wireless hubs. As you can see, the higher the injection rate is, the better our proposed topology and routing are. When the reliability of wireless hubs is high, the benefit of the fault-tolerant routing is not impressive. On the other hand, in state-of-the-art HWNoC-based routings, when there is a fault in a wireless hub, the hub will be unreachable and the related packets will use basic routing (here XY routing) to reach their destination. Hence, the efficiency of the wireless channel in the network will be reduced, which leads to increasing the average latency and power consumption of the whole network.

In Fig. 7(b, c), the wireless hub reliability is assumed to be 90% and 80% respectively. As it can be seen, the difference in these charts is greater and the improvement of our proposed architecture is evident. In the Fig. 7(c), the reliability is considered 80% and the difference between the charts is still higher. In fact, the higher the likelihood of network fault and the higher the injection rate, the better the architecture provided than other topology combinations and routing in other reported works. Meanwhile, the higher the error rate, the more important routing becomes. If we check the diagram of 7(c), then it can be seen that when we combine the proposed routing algorithms with the topology presented in [24], it results in a better improvement than the proposed topology combination in [24] with routing [22]. Table 4 lists the summary of average delay improvements provided by our proposed architecture compared to other architectures.

Figure 8 shows the decrease in delay at the injection rate of 0.003 when the wireless hub reliability increases. In fact, this figure shows the flexibility of the various HWNoC architectures. Note that in the injection rate of 0.003, there are enough packets in the network while the network is still not saturated. As can be seen, our proposed architecture performs better than its counterparts.

As shown in Fig. 8, the proposed architecture, which combines topology and proposed routing, has the best performance. It has already been said that the greater the reliability of wireless hubs, the more important the topology will be than routing. If the reliability is reduced to less than 60%, the architecture that is derived from the combination of topology [24] and proposed routing in this article performs better than the proposed topology architecture in this paper and routing [22]. For a topology with 256 nodes and 8 hubs, at the injection rate of 0.003 packets/node/cycle, the proposed architecture has decreased 27.93%, 13.29%, and 12.14% of the average delay of Architectures No. 1, 2, and 3 respectively.

5 Conclusions

Since HWNoC-based systems are more error-prone than the conventional ones, in this paper, we proposed a new topology along with a novel routing algorithm to tolerate the failure of the wireless hubs. In the proposed approach,

Table 4 Average percentage of delay improvements provided by architecture No. 4 of Table 3 compared to the other architectures

Architecture			Reliability of wireless hubs		
No.	Routing algorithm	Topology	1 (%)	0.9 (%)	0.8 (%)
1	Rezaei et al. [24]	Hu et al. [22]	21.85	36.96	39.94
2	Current work	Hu et al. [22]	6.86	11.54	19.27
3	Rezaei et al. [24]	Current work	19.92	34.78	35.62

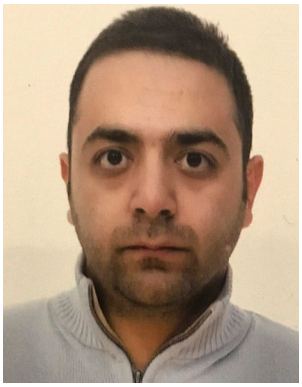
once a wireless hub becomes faulty, the best alternative hub will be chosen and all the packets that have high average hop-count will be routed through this alternative hub. In comparison with the state-of-the-art works, the proposed architecture shows significant improvements in terms of robustness, congestion management, and resilience.

References

1. Dally, W., & Towles, B. (2001). Route packets, not wires: On-chip interconnection networks. In *Proceedings of the design automation conference* (pp. 684–689), CA, USA.
2. Rezaei, A., Zhao, D., Daneshtalab, M., & Zhou, H. (2017). Multi-objective task mapping approach for wireless NoC in dark silicon age. In *International conference on parallel, distributed and network-based processing (PDP)* (pp. 589–592), St. Petersburg, Russia.
3. Chen, Y. Y., Chang, E. J., Hsin, H. K., Chen, K., & Wu, A. (2017). Path-diversity-aware fault-tolerant routing algorithm for network-on-chip. *IEEE Transactions on Parallel and Distributed Systems, 28*(3), 838–849.
4. Akbar, R., & Safaei, F. (2018). A novel adaptive congestion-aware and load-balanced routing algorithm in networks-on-chip. *Computers & Electrical Engineering, 71,* 60–76.
5. Pavlidis, V. F., & Friedman, E. G. (2007). 3-D topologies for networks-on-chip. *IEEE Transactions on Very Large Scale Integration (VLSI) Systems, 15*(10), 1081–1090.
6. Feero, B. S., & Pande, P. P. (2009). Networks-on-chip in a threedimensional environment: A performance evaluation. *IEEE Transactions on Computer, 58*(1), 32–45.
7. Rezaei, S. H. S., Modarressi, M., Aminabadi, R. Y., & Daneshtalab, M. (2016). Fault-tolerant 3-D network-on-chip design using dynamic link sharing. In *Design, automation & test in Europe conference & exhibition (DATE)* (pp. 1195–1200), Dresden, Germany.
8. Salamat, R., Khayambashi, M., Ebrahimi, M., & Bagherzadeh, N. (2018). LEAD: An adaptive 3D-NoC routing algorithm with queuing-theory based analytical verification. *IEEE Transactions on Computers, 67*(8), 1153–1166.
9. Shacham, A., Bergman, K., & Carloni, L. P. (2008). Photonic networks-on-chip for future generations of chip multiprocessors. *IEEE Transactions on Computers, 57*(9), 1246–1260.
10. Pan, Y., Kumar, P., Kim, J., Memik, G., Zhang, Y., & Choudhary, A. (2009). Firefly: Illuminating future network-on-chip with nanophotonics. In *Proceedings of the annual international symposium on computer architecture* (pp. 429–440), Texas, USA.
11. Chang, M. F., Cong, J., Kaplan, A., Naik, M., Reinman, G., Socher, E., et al. (2008). CMP network-on-chip overlaid with multi-band RF-interconnect. In *High performance computer architecture (HPCA)* (pp. 191–202), Salt Lake City, UT, USA.
12. Ganguly, A., Chang, K., Deb, S., Pande, P. P., Belzer, B., & Teuscher, C. (2011). Scalable hybrid wireless network-on-chip architectures for multicore systems. *IEEE Transactions on Computers, 60*(10), 1485–1502.
13. Deb, S., Ganguly, A., Pande, P. P., Belzer, B., & Heo, D. (2012). Wireless NoC as interconnection backbone for multicore chips: Promises and challenges. *IEEE Journal on Emerging and Selected Topics in Circuits and Systems, 2*(2), 228–239.
14. Deb, S., Chang, K., Yu, X., Sah, S. P., Cosic, M., Ganguly, A., et al. (2013). Design of an energy-efficient CMOS-compatible NoC architecture with millimeter-wave wireless interconnects. *IEEE Transactions on Computers, 62*(12), 2382–2396.
15. Chang, K., Deb, S., Ganguly, A., Yu, X., Sah, S. P., Pande, P. P., et al. (2012). Performance evaluation and design trade-offs for wireless network-on-chip architectures. *ACM Journal on Emerging Technologies in Computing Systems (JETC), 8*(3), 23.
16. Wettin, P., Murray, J., Kim, R., Yu, X., Pande, P. P., & Heo, D. (2014). Performance evaluation of wireless NoCs in presence of irregular network routing strategies. In *Design, automation and test in Europe conference and exhibition (DATE)* (pp. 1–6), Dresden, Germany.
17. Dehghani, A., & Jamshidi, K. (2016). A novel approach to optimize fault-tolerant hybrid wireless network-on-chip architectures. *ACM Journal on Emerging Technologies in Computing Systems (JETC), 12*(4), 45.
18. Dehghani, A., & Jamshidi, K. (2015). A fault-tolerant hierarchical hybrid mesh-based wireless network-on-chip architecture for multicore platforms. *The Journal of Supercomputing, 71*(8), 3116–3148.
19. Ogras, U. Y., & Marculescu, R. (2006). It's a small world after all: NoC performance optimization via long-range link insertion. *IEEE Transactions on Very Large Scale Integration (VLSI) Systems, 14*(7), 693–706.
20. Bahrami, B., Jamali, M. A. J., & Saeidi, S. (2017). A hierarchical architecture based on traveling salesman problem for hybrid wireless network-on-chip. *Wireless Networks*, *1*, 1–14.
21. Wang, C., Hu, W. H., & Bagherzadeh, N. (2012). A load-balanced congestion-aware wireless network-on-chip design for multi-core platforms. *Microprocessors and Microsystems, 36*(7), 555–570.
22. Hu, H. W., Wang, C., & Bagherzadeh, N. (2015). Design and analysis of a mesh-based wireless network-on-chip. *The Journal of Supercomputing, 71*(8), 2830–2846.
23. Rezaei, A., Safaei, F., Daneshtalab, M., & Tenhunen, H. (2014). HiWA: A hierarchical wireless network-on-chip architecture. In: Conference on high performance computing & simulation (HPCS) (pp. 499–505), Bologna, Italy.
24. Rezaei, A., Daneshtalab, M., Safaei, F., & Zhao, D. (2016). Hierarchical approach for hybrid wireless network-on-chip in many-core era. *Computers & Electrical Engineering, 51,* 225–234.
25. Rezaei, A., Daneshtalab, M., & Zhao, D. (2017). CAP-W: Congestion-aware platform for wireless-based network-on-chip in many-core era. *Microprocessors and Microsystems, 52,* 23–33.
26. Zhao, D., & Wang, Y. (2008). SD-MAC: Design and synthesis of a hardware-efficient collision-free QoS-aware MAC protocol for wireless network-on-chip. *IEEE Transactions on Computers, 57*(9), 1230–1245.
27. Afsharmazayejani, R., Yazdanpanah, F., Rezaei, A., Alaei, M., & Daneshtalab, M. (2018). HoneyWiN: Novel honeycomb-based wireless NoC architecture in many-core era. In *International symposium on applied reconfigurable computing (ARC)* (pp. 304–316).
28. Kumar, A., Peh, L. S., & Jha, N. K. (2008). Token flow control. In *Proceedings of the international symposium on microarchitecture* (pp. 342–353).
29. Palesi, M., Collotta, M., Mineo, A., & Catania, V. (2015). An efficient radio access control mechanism for wireless network-on-chip architectures. *Journal of Low Power Electronics and Applications, 5*(2), 38–56.
30. Ganguly, A., Wettin, P., Chang, K., & Pande, P. (2011). Complex network inspired fault-tolerant NoC architectures with wireless links. In *International symposium on networks on chip (NoCS)* (pp. 169–176), Pittsburgh, Pennsylvania.
31. Catania, V., Mineo, A., Monteleone, S., Palesi, M., & Patti, D. (2016). Cycle-accurate network on chip simulation with noxim.

ACM Transactions on Modeling and Computer Simulation (TOMACS), 27(1), 4.

Seyed Hassan Mortazavi received the B.Sc. degree in computer software engineering from Islamic Azad University, Najaf Abad, Isfahan, Iran, in 2014 and the M.Sc. degree from the Shahid Beheshti University, Tehran, Iran, in 2017. His research interests focus on design and test of large SoCs, with particular emphasis on their data communication infrastructures and cloud base infrastructures for IoT systems. He has practical experiences on restful cloud base software systems on unix base platforms as Back-end developing.



Reza Akbar received the B.Sc. degree in Computer Hardware Engineering from Khaje Nasir Toosi University, Tehran, Iran, in 2010, and M.Sc. degree in Computer Architecture Engineering from Sharif University, Tehran, Iran, in 2012. He is currently Ph.D. student in Computer Architecture Engineering in Department of Computer Science and Engineering, Shahid Beheshti University, Tehran, Iran. Currently, his researches focus on Networks-on-Chip, High Performance Computer Systems, and Complex Networks.

Farshad Safaei received the B.Sc., M.Sc., and Ph.D. degrees in Computer Engineering from Iran University of Science and Technology (IUST) in 1994, 1997 and 2007, respectively. He is currently an assistant professor in the Department of Computer Science and Engineering, Shahid Beheshti University, Tehran, Iran. His research interests are Performance Evaluation of Computer Systems, Networks-on-Chips, and Complex Networks.

Amin Rezaei is currently a Ph.D. candidate in Electrical Engineering and Computer Science at Northwestern University, the USA. He received his B.Sc. degree in Computer Engineering from University of Isfahan, Iran, in 2011 and two M.Sc. degrees one in Computer Engineering from Shahid Beheshti University, Iran, in 2014 and the other in Computer Science from University of Louisiana at Lafayette, the USA, in 2016. His main research interests include Parallel Computing, Dark Silicon, and Logic Encryption.