IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received December 2, 2019, accepted January 17, 2020, date of publication January 29, 2020, date of current version February 5, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2970215

A Novel Low-Latency Regional Fault-Aware Fault-Tolerant Routing Algorithm for Wireless NoC

YIMING OUYANG[1], **QI WANG**[2], **MENGXUAN RU**[1], **HUAGUO LIANG**[2], **AND JIANHUA LI**[1]

[1]School of Computer and Information, Hefei University of Technology, Hefei 230009, China
[2]School of Electronic Science and Applied Physics, Hefei University of Technology, Hefei 230009, China

Corresponding author: Qi Wang (keywenchester@outlook.com)

This work was supported in part by the National Natural Science Foundation of China (NSFC) Research Projects under Grant 61874157, Grant 61674048, and Grant 61874156, and in part by the Natural Science Foundation of Anhui Province China under Grant KJ2018A0783.

ABSTRACT WiNoC has become a promising on-chip interconnect architecture. Due to the integration and manufacturing limits of wireless interconnects in nanotechnology, WiNoC systems are more susceptible to high failure rates. In this paper, we propose a novel fault-tolerant routing algorithm based on regional fault-aware techniques for link failures in WiNoC. We trade off performance and overhead by adopting 2-hop awareness for wireless nodes and 1-hop awareness for wired nodes. And the node status is not defined merely based on its own faulty condition but with a ‘‘double sensing’’ mechanism. Besides, since congestion is prone to occur around faulty nodes, we consider congestion mitigation in fault-tolerant algorithm design. Simulation results demonstrate that compared with several counterparts, the performance benefits clearly overweigh the overhead with acceptable area increase. The newly proposed routing algorithm significantly enhances the robustness of the system.

INDEX TERMS Wireless network-on-chip, fault-tolerant routing, fault-aware mechanism.

I. INTRODUCTION

As silicon technology continues to advance, large-scale chip multiprocessors (CMPs) and system-on-chip (SoCs) have become mainstream in design. Network-on-chip (NoC) technology can integrate various processors and memory unit into a single chip, which has become a leading-edge communication architecture in chip design [1]. Although NoC has its unique advantages, as the number of processors increases, integrated circuits (ICs) architecture becomes more and more complex. One major performance limitation of traditional NoC comes from multi-hop communication of planar metal interconnects. Data transmission between distant nodes leads to high latency and power consumption [2]. According to the International Technology Roadmap for Semiconductors (ITRS) [3], metal wires will no longer meet on-chip system performance requirements, and new interconnect modes need to be introduced eagerly. Researchers have explored this subject in several direction such as 3DNoC, Optical NoC, Radio Frequency NoC (RFNoC) [4]–[6]. These methods contribute in reducing latency and power consumption of the network on the chip to a certain extent, but they still have limits in interconnection. 3DNoC must overcome technical and manufacturing challenges to achieve mass production. Although optical interconnects have the advantages of high throughput and low transmission power consumption, they still face problems such as manufacturing cost, temperature sensitivity of photonic devices, and design complexity [5]. Furthermore, although RF-interconnects (RF-I) can be built using complementary metal oxide semiconductor (CMOS) technology, they require long on-chip transmission lines without eliminating any existing links, which can lead to wiring difficulty and large area overhead [6].

Wireless Network-on-chip (WiNoC) can achieve efficient long-distance communication with advantage of being compatible with CMOS wireless technology. It is considered by major researchers to be an alternative scheme to traditional NoC interconnect architecture [7]. Wireless channels do not eliminate wired communications in WiNoC; instead, they are considered as a complement to wired communications [8]. [9] shows that the hybrid layered wireless architecture is beneficial to achieve a good tradeoff between delay and energy consumption under limited area resources. As technology expands into the Nano-domain, transistor

The associate editor coordinating the review of this manuscript and approving it for publication was Liang Yang.

size shrinkage, power voltage reduction, and operating frequency increase seriously affect the reliability of integrated circuits (CMOS VLSI) [10]. Although hybrid WiNoC has the advantages of wireless communication, wireless technology is prone to various failures due to high integration density and high complexity of system circuits [11], [12]. A wired link failure in WiNoC can cause many system level failures such as packet replication, drop, misroute, and transmission delay [13]. Wireless communication components require high sensitivity to noise sources, and manufacturing problems increase the failure rate of wireless technologies [11], [14]. In the hybrid WiNoC architecture, the wireless channel is responsible for long-distance transmission and plays a significant role in enhancing system performance. When WIs are in fault, they can't function well as shortcut in router path, leading to degradation of the overall network performance [15]. Therefore, fault tolerance has become a major consideration in current and future WiNoC architecture design. The emergence of fault-tolerant design adds a powerful protective shield to the entire system on chip. At the network level, a routing algorithm able to tolerate link failures is beneficial to increasing the stability of the WiNoC system significantly [12].

In WiNoC, WR is more prone to be congested than traditional NoC since it is shared among all cores (PE) [16]. Excessive transmission of packets routed to the WR results in reduced network throughput and high latency [9]. In addition, congestion is more likely to occur around the faulty router due to inherent characteristics of failures on the chip, especially when the wireless link fails. The number of wireless channels in WiNoC is limited hence these valuable resources must be effectively utilized. So far, researches have proposed a variety of methods to prevent WR congestion. Some of these methods solve this problem by adding buffers and proposing congestion-aware routing algorithms that can be used to significantly reduce latency in WiNoC when adopting WR placement algorithms or application mapping techniques [17]. However, few of them have suggested WiNoC congestion solutions when wireless links fail.

In this paper, an adaptive fault-tolerant routing algorithm for Regional Fault Awareness (RFA) is proposed to address permanent link failure in WiNoC. The algorithm can subtly combine fault information with congestion information in WiNoC, which provides a good view for packet routing, reduces the probability of encountering a faulty link during routing, and selects the fittest path for packet routing. The additional congestion awareness scheme balances the network load in one faulty network. The major contributions of this paper are as follows:

(1) Proposes a novel method for transmitting node fault status in the form of information aggregation. Routers can sense the fault level around themselves, avoid highly faulty area in RC phase, reduce latency and prevent backtracking.
(2) Adopts a compromised 2-hop information transmission mechanism. The congestion/failure information of the wired node is adjacently perceived, and the congestion/fault information of the wireless node is sensed by two hops. Thereby reducing the wiring cost;
(3) The scope of fault and congestion awareness is expanded so the packets are not congested in wireless routers and its surrounding routers.

The rest of the paper is organized as follows: Section 2 summarizes some related work in fault tolerance and congestion control for wired and wireless NoC in recent years. Section 3 describes the WiNoC infrastructure and WiNoC congestion and failure analysis. Section 4 elaborates the proposed RFA routing algorithm. Section 5 is the experimental analysis. Finally, Section 6 concludes this paper.

II. BACKGROUND AND RELATED WORKS

Faults in NoC include permanent and transient failures and each one can occur for a variety of reasons [10]. Soft errors, crosstalk, and voltage-induced delay errors often cause transient faults in NoC systems, and related studies generally use ECC encoding [18]. Due to the inherent characteristics of the wireless channel and the electromagnetic interference of the wireless system and other components on the chip, the transient failure probability in the WiNoC architecture is higher than that of the conventional NoC architecture. [18] proposes a unified error control code (ECC) framework for hybrid WiNoC architecture. This solution increases system reliability but creates additional area overhead. The reliability of the mixed WiNoC architecture based on Code Division Multiple Access (CDMA) is explored in [19], which proves that WiNoC can achieve reliable communication using CDMA-based wireless interconnection. Due to the severe challenges of wireless interface design and integrated environments in nanoscale domains, the probability of permanent failures is also increasing [19]. Manufacturing defects, hardware aging, thermal stress, and physical stress can cause permanent failures such as logic shorts [10]. Permanent failure can cause irreversible damage to the circuit. Whereas permanent failures are not as common and frequent as transient failures on a chip, when a component fails, it is impossible to repair or replace it on the chip. In this case, it is necessary to reroute the packets on alternative paths so that the communication infrastructure remains intact [20]. Because the wireless interconnection method is still in its early stages, there will be problems with manufacturing process defects [11]. There are few studies on permanent failures in the WiNoC architecture. The authors of [11], [15] proposed a performance level based on the complex network WiNoC architecture in the presence of wireless link failures. They use the power law model to establish wireless links between routers. This approach enhances NoC performance while minimizing performance degradation in the event of wireless links failure. In [14], different trade-offs for the layered hybrid WiNoC architecture are proposed. It is pointed out that although the small world WiNoC has lower performance than the layered WiNoC, it can provide higher fault tolerance and tolerate wireless link failure. In [20], a method for solving the fault-tolerant wireless router placement problem in Mesh networks is proposed. Based on this, an efficient fault-tolerant communication protocol is proposed. In the case of a WR fault, the alternative WR is selected for wireless transmission. In these studies, either the wireless link failure problem has not been completely solved, or the alternate path is selected by adding other WRs, which increases the system area and power consumption overhead. It also does not cover issues such as hot spots around the fault and unbalanced network load.

Existing fault-tolerant routing algorithms may recommend unnecessary paths and tend to generate hot spots around faulty nodes. Therefore, it is necessary to introduce congestion awareness in the fault tolerance mechanism. Relative to the congestion problem of WiNoC, many researchers have proposed related research. Throttling-based congestion control for wired links is one of the methods for solving congestion problems in traditional NoC [21]. In [22], the authors used a source throttling-based congestion control mechanism with application-level awareness that controls the injection rate. For wireless links in [23], the authors proposed a router architecture that distributes traffic around a wireless router through a crossbar switch and proposes a WiNoC congestion control routing algorithm to mitigate congestion. In [17], a balancing parameter is added to the routing algorithm to control the congestion of WR. By increasing the network size and channel utilization, the value of the equalization parameter increases exponentially. However, under higher traffic load, the value of the equalization parameter increases, and the routing algorithm may not select the wireless link to transmit the packet, resulting in an increase in delay.

Limits in existing methodology: 1. The congestion occurred around a faulty link is not yet settled; 2. The proposed routing algorithm can't make correct routing decisions according to the fault region of the network. It is possible to select non-minimum routing strategy and route the packets to the area with high failure rate. In order to improve the robustness of WiNoC, maintain high performance and fault tolerant ability under faults, two key issues need to be covered: (A) avoiding blocking paths routing and balance the load; (B) When facing permanent failure, the fault can be tolerated and the system performance can be maintained on an acceptable level. In this paper, for link failures in WiNoC, an adaptive routing strategy combining fault tolerance and congestion awareness is studied to maintain the graceful degradation of WiNoC performance.

III. WiNoC ARCHITECTURE AND FAULT MODEL

WiNoC establishes a wireless channel through on-chip wireless communication technology. [24] pointed out that WiNoC with ultra-wideband (UWB) technology uses Meander type dipole antennas with transmission range and data rate of 1 mm and 1.16 Gbps, respectively. But UWB interconnected WiNoC cost much more area overhead, and require more multi-hop when long-distance data transmission. [25] confirmed that a wireless interface based on a Zigzag antenna and a millimeter-wave transceiver can be implemented on-chip, and its transmission range, data rate and energy consumption per bit of data are 20 mm, 10-100 Gbps, and 2.3 pJ, respectively. It is suitable for establishing high-speed channels for long-distance data transmission in WiNoC and has the advantages of low delay and low power consumption. [26] pointed out that the antenna design based on the multi-walled carbon nanotube (MWCNT) process can push the WiNoC data transmission frequency to THz, and its transmission range, data rate, and energy consumption per bit of data are 23 mm, 240 Gbps, and 0.33 pJ, respectively. However, due to the technical bottleneck and reliability of CNT antenna size integration, implementation in WiNoC design is severely limited.

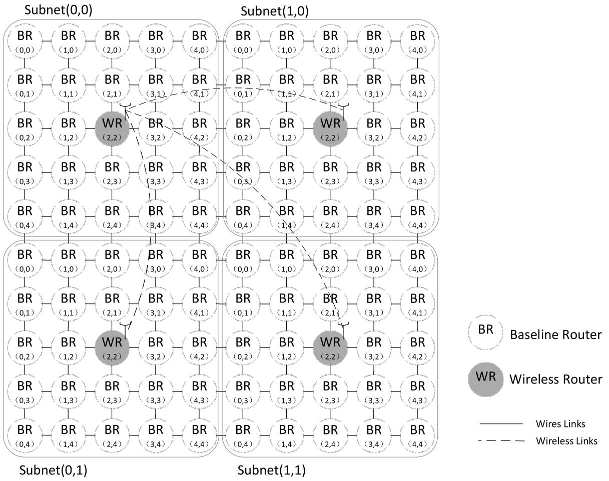


FIGURE 1. WiNoC topology.

The method proposed in this paper can be used in different architectures of WiNoC. Due to the versatility and applicability of 2D-mesh topology, this paper has been studied on a network with a hybrid wired/wireless architecture. Fig.1 shows a 10×10 layered hybrid WiNoC. The network is divided into four 5×5 subnets. Each subnet has a wireless router (WR) which implements with wireless interface (WI) responsible for remote data communication at the center of the subnet. Low-level communication is performed over wired links, and higher levels are communicated through wireless channels. Wireless routers function as gateways in remote communications between cores. This paper uses Zigzag on-chip antenna and millimeter wave transceiver, which has the characteristics of long transmission distance, high gain and strong anti-interference [27].

A. FAULT MODEL AND FAULT DETECTION

In this paper, we mainly study the permanent link failure that may occur in the WiNoC system, and the fault is regarded as a completely disconnected link. A link failure is considered to be unidirectional. A router failure means all its output links fail. In each router, we use a eight-bit fault vector to denote the fault status of its four output links, another eight vector to represent the fault status of the four input links.

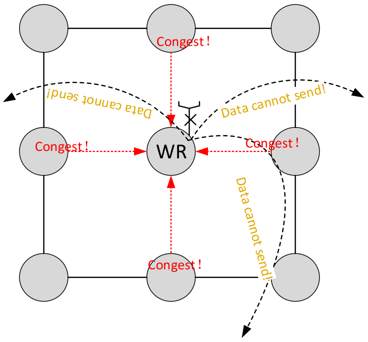


FIGURE 2. WR failure causes congestion in surrounding routers.

When the link fails, it stops providing service to the network and the packet is only transmitted over the faultless link. For wireless transceivers, failures at any stage of the Power Amplifier (PA), Serializer, and Antenna Transmitter may cause the WI to fail and disabled to send packets(This article takes the approach of disabling the wireless transceiver to simulate the failure of the wireless transmitter). The WI failure in WR is that the router cannot perform wireless transmission, and its wired router is the same as other wired routers. Permanent fault detection can be detected by BIST [28] or by sending a detection vector [29], which was done in our previous work. After detecting the faulty link by an appropriate detection mechanism such as the built-in self-test (BIST) mechanism, the configuration register can be updated during the reconfiguration process. The failure detection, notification, and update of the corresponding configuration bits of the router occur before the normal operation begins. Moreover, we assume that the router can automatically resume its operational processing after completing the reconfiguration.

B. CONGESTION ANALYSIS

Congestion can occur around the faulty node in the NoC and cause adverse effect on system performance, increasing latency and lowing throughput [35]. At present, according to the routing algorithm related to WiNoC, many packets are selected to be forwarded over the wireless link [23]. Therefore, the wireless router bears a large part of the traffic in the network, and many scholars have improved this to solve the congestion problem of the wireless router. However, they did not consider the scenario when WI failed, as shown in Fig.2. Once the WI fails, many incoming wireless packets will not be processed. The previous method does not apply here. When the WR stops to process a large number of arriving packets, it will cause more serious congestion in the WR and its neighboring routers, resulting in a decrease in transmission efficiency and seriously affecting the system performance of the on-chip network [20].

C. FAILURE AND CONGESTION AWARENESS

A faulty router will be reconfigured and inform all its neighbor nodes of their faulty status. For instance, they disable the output port, simultaneously, the downstream router disables the corresponding input port. Each router stores the neighbor routers' and its own status in the fault status register. This approach is similar to the work proposed in [30]. We assume that fault detection is triggered, notified and updated with the appropriate configuration bits(register) before the router acts normal operation.

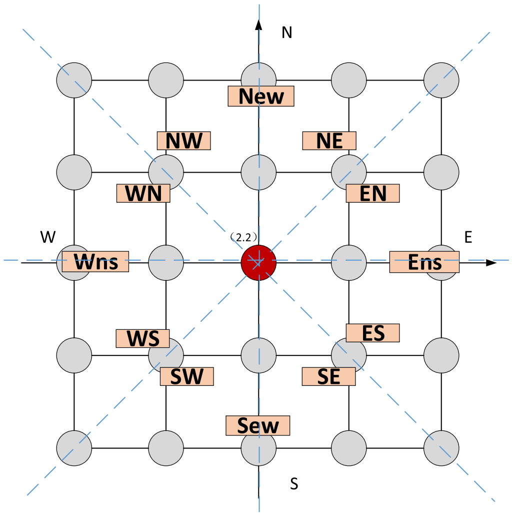


FIGURE 3. Priority-based zoning.

In this work we trade off area overhead and performance in priority, setting wired node to sense information in one jump distance and wireless node sense in two jumps. In adaptive routing, two-hop aware of congestion information and fault information is a common method in NoC. By transmitting the congestion and fault information of all nodes in two hops and then performing route calculation, the network performance can be greatly improved, but a large amount of wiring and registers are required, cost more area consumption on chip. This paper proposes a compromise approach that does not use more wiring and achieves performance improvements by combining routing algorithms. In this paper, wired nodes transfer status information within adjacent nodes. While by adopting regional fault information exchange, wired nodes will be able to sense surrounding nodes status more than one hop, which fill the gap of one-hop information exchange. Wireless nodes can sense information within two hops in an effort to mitigate congestion occurring around faulty wireless routers. In WiNoC, WI only takes a small part in whole system, which means two-hop awareness is not used in high frequency. This mechanism reduces wiring cost and area consumption compared to adopting two-hop awareness in all nodes, achieves better promise in cost and performance.

IV. LOW-LATENCY CONGESTION-AWARE FAULT-TOLERANT ALGORITHM

To simplify the relatively position between destination node and current node, we assume that the destination node is in four major direction from current node (New,Ens,Sew,Wns). In order to optimize the selection of the output port and provide a better adaptability to the routing algorithm, we propose a priority-based region partitioning as shown in Fig.3

(i.e., In the New direction, priority: N > E = W). By comparing the placement of the current node and the destination node, select the direction with largest Manhattan distance and set this direction as first priority. For example, for node (2, 2), if the destination node is (5, 6) in the NE area, the priority in the N direction is greater than the E direction; when the destination node is on the dotted line, the two related directions can be used as the preferred port at the same time, and the priority is the same. That is, if the destination node is (5, 5), both the N and E directions can be used as the preferred port. The specific selection of the current port can be combined with the failure and congestion information of the adjacent node links.

In this paper, a fault-tolerant routing algorithm based on regional fault perception is proposed for WiNoC. Because congestion is dynamic in the network, it is only necessary to sense the congestion of neighboring nodes. For tolerant faulty links, this paper proposes a new fault-aware method to decide the level of failure status around the current nodes by transmitting surrounding network failure information. By selecting a direction with a lower degree of failure, we reduce the probability of encountering a failure in the routing process. Besides, we also consider the congestion as the other matric when choose the routing direction.

A. WIRED LINK REGIONAL FAULT AWARENESS

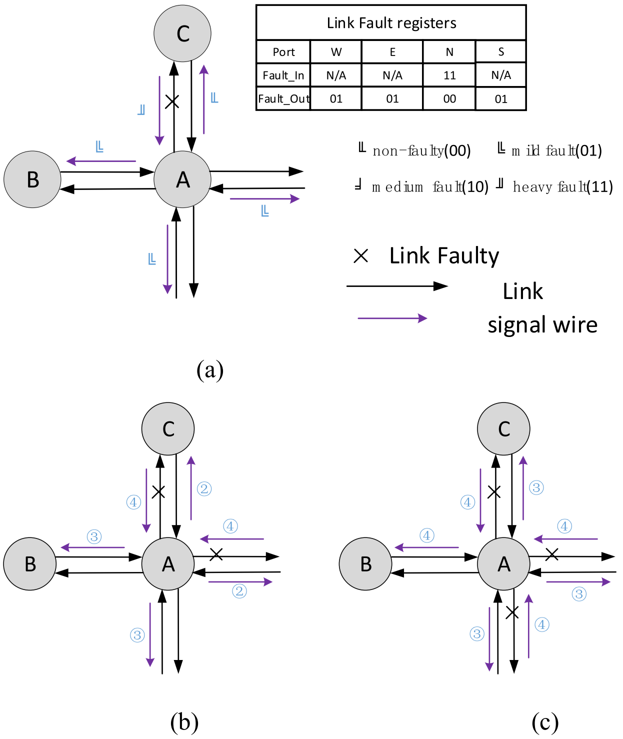
For failures in wired link, we double sense the conditions around a certain WI to define the extent of fault region, so the fault level can be measured more precisely.

Node status are not defined as faulty or not fault purely based on its own conditions, but with considering direction as a metric. One node may be heavily faulty to nodes in a certain direction but fault-free to another direction. Initially, the link information passed between nodes is set to no fault by default. That is, the link fault register records that the fault information (Fault_in) of the surrounding nodes is N/A until the fault information of the surrounding nodes is received, then register is updated. And when we say node incurs failures in certain output links, we indicate that it's fault-free in all other links. The node status is defined as follows:

Definition 1: When current node has no output link failure, the node is defined as a non-faulty node in all four directions;

Definition 2: When current node A has one faulty output link, then for nodes in this output direction, A is fault-free and for nodes in other three directions A is slightly fault. As shown in Fig.4a, current node A has one output link failure in N direction and there is no link failure in the W, E, and S directions. Then A is fault-free for Node C, passing semaphore 00 and updating the table. A is slightly fault for nodes in other three directions, giving semaphore 01 and updating Fault_out table;

Definition 3: When current node A has two faulty output links referring to a certain direction, as shown in Fig.4b, Node A is moderately fault to B (setting semaphore 10) and slightly fault to C (setting semaphore 01);



Link Fault registers				
Port	W	E	N	S
Fault_In	N/A	N/A	11	N/A
Fault_Out	01	01	00	01

(a)

(b)

(c)

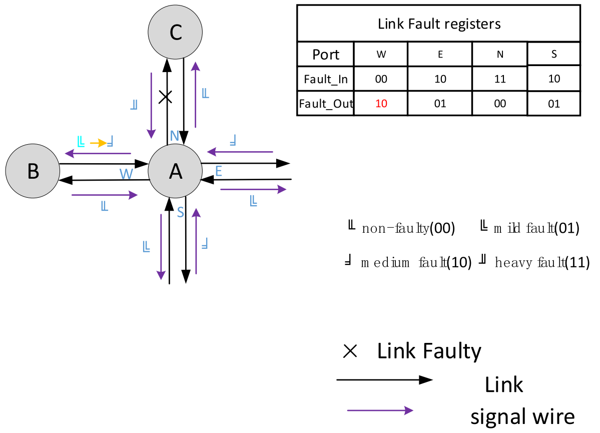
FIGURE 4. Node state definitions.

Definition 4: When current node has three output link faults, as shown in Fig.4c, the output link of the A-node in N, E, and S directions fails. Define A as a severely fault node to nodes in W direction, transmitting semaphore 11, indicating that this direction is not routable. Defining A node as a moderate fault node to the other three directions, transmitting semaphore 10.v

When the transmission direction of link fails, the input/output fault register in this direction always records link status as fault (viz., status information from router in a fault direction is not accepted to update register). As shown in Fig. 4a, the link in the N direction is faulty, and data cannot be transmitted to the N direction, that is, the information of the direct update Fault_in is 11, indicating that the N direction is unavailable.

The transmission of fault information between routers: the node transmits the fault information to the neighboring nodes, so that the current node senses the fault state of the surrounding nodes and updates the Fault_in table in Fault register. Redefining the state of the faulty node: first, the node feeds back the fault information to the surrounding nodes according to its fault state, and accepts the fault information of the surrounding nodes; Then the integrated node feeds back the fault state information of the node according to the fault state of the surrounding node, so as to achieve the effect of reginal fault aware.

After a node perceives surrounding links and nodes status from the other three directions (Fault_in record), its status is redefined and feeds back the new status to nodes in the left direction. Redefine rules are shown as follow:



Link Fault registers				
Port	W	E	N	S
Fault_In	00	10	11	10
Fault_Out	10	01	00	01

FIGURE 5. Fault status updating and fault register table.

(1) Node A is redefined as faulty-free(①) to a certain direction.

case1: A receive no fault(1∗①) from other three direction;

case2: A receives one or two mild faults(2∗②);

case3: A receives one moderate fault (1∗③);

case4: A receives one moderate fault (1∗③) and one mild fault(1∗②);

case5: A receives three mild faults (3∗②);

case6: A receives two moderate faults(2∗③)

case7: A receives one mild fault and one moderate fault (1∗② + 1∗③);

(2)Node A is redefined as a mild fault(②) node to a certain direction.

case1: A receives a mild fault and two moderate faults (1∗ + 2∗③);

case2: A receives three moderate faults(3∗③);

case3: A receives one mild fault (②) and one faulty link(1∗② + 1∗④);

case4: A receives one medium fault and one faulty link (1∗③ + 1∗④);

case5: A receives two mild fault and one faulty link (2∗② + 1∗④);

case6: A receives one mild fault, one moderate fault and one faulty link(1∗② +1∗③ + 1∗④);

(3)Node A is redefined as a moderate fault(③).

case1: A receives two moderate faults and one link fault(2∗③ + 1∗④);

case2: A receives one mild fault and two faulty links(1∗② + 2∗④);

case3: A receives one moderate fault(③) and two faulty links(1∗③ + 2∗④);

(4) Node A is redefined as heavily faulty(④) only when A receives three faulty links(④) from other three direction.

After receiving surrounding fault information, the node resends the fault status information to the surrounding nodes if fault level rises. Fig.5 illustrates how node status changes after fault information get updated. Example 1, the node sends a mild fault information(②) to W direction. After receiving fault information from surrounding nodes, the faults around it are a severe fault (④) in the N direction and a moderate fault (③) in E direction and S direction. Therefore, it is necessary to redefine the node state, upgrade from a mild fault to a moderate fault(② −>③), update the Fault_out to 10 in W direction, and then send a moderate fault information(③) to W direction. In example 2, node sends the fault-free information(①) to N direction. After receiving fault information, the faults around it are fault-free(①) in the W direction and moderate fault(③) in E direction, moderate fault(③) in the direction. Therefore, the node status is no fault(①) to N direction, and the fault information is transmitted unchanged in the N direction. Example 3: This node sends a mild fault (②) to the E direction. After receiving fault information transmitted by the surrounding nodes, faults around it are severe faults(④) in the N direction, and the W direction is faultless(①), the S direction is moderately faulty(③), so there is no need to redefine node state, maintaining mild faulty to E direction.

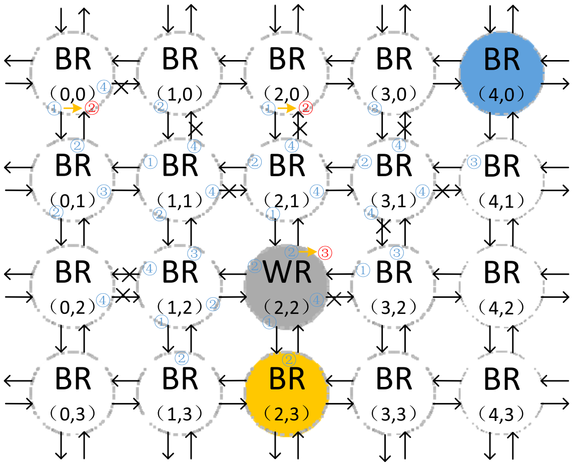


FIGURE 6. Updating fault status information between nodes.

Fig.6 shows the fault status information updating between nodes (updated one is marked as red). Because moderate faults > mild faults > no faults, the definition of various fault states has their thresholds, when updating fault information and delivery, it will only cover the faults, so it will not affect the whole network. For example, the S port of BR(0,0), the S port of BR(2,0), and the N port of WR(2,2) in the figure are updated after receiving the information.

The routing algorithm can optimize the path selection according to the degree of fault state around the node. Through the sensing of the fault area, the fault area can be effectively avoided, and the data packet can be routed to the area with no fault or less fault.

B. WIRELESS LINK FAILURE AND CONGESTION AWARENESS

As shown in Fig.7, this paper expands the state sensing of wireless node failure and congestion to two-hop awareness, so that nodes in two-hop distance can perform dynamic routing when the wireless interface of the wireless node fails. This relieves surrounding nodes from congestion when the wireless interface fails, balancing the wireless node and the network load around it. Two-hop awareness can help selecting

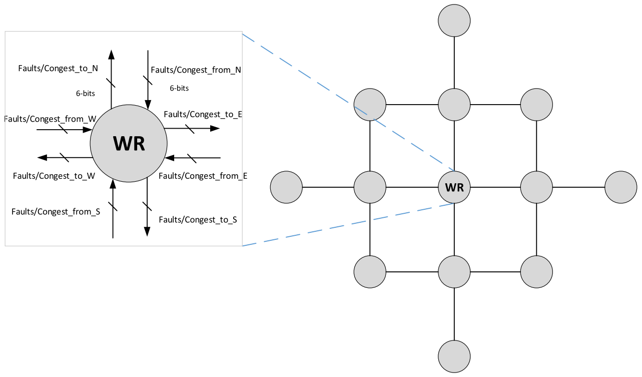


FIGURE 7. WR two-hop awareness.

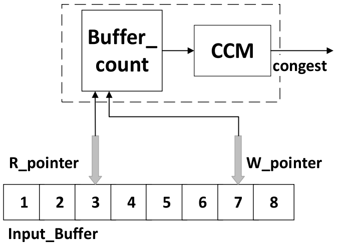


FIGURE 8. Congestion counter.

routes for routing decisions. Nodes out of two-hop distance set fault-free and no congestion as wireless router status and only when packets touch the sensing range, dynamic routing based on two-hop sensing functions to decide if wired link is more appropriate.

C. CONGESTION PRIORITY AWARENESS

Wireless links accommodate single hop shortcuts for long-distance nodes, while this convenience takes adverse effects on global traffic flow distribution, leading the non-uniform distribution even worse, so to cause congestion. To address this problem, we fuse congestion aware design with the fault-tolerant algorithm. NoC provides abundant connection resources, high bit width and working frequency, which can meet the bandwidth requirements well. The main bottleneck is not the link bandwidth, but the cache capacity. For NoC congestion metrics, we use metrics many researchers are using the buffer capacity carried out. As shown in Fig.8, the flit stored in the buffer is counted by the congestion counter, and t buffer occupancy of the buffer is divided into four levels, 0% idle (00), 50% occupied (01), 75% busy (10) and 100% congestion (11). Perceived by congestion priority, it notifies the nodes within two hops. In Fig.8, the CCM is responsible for the management and delivery of congestion signals. Congestion metrics can only be used as an adjunct to path selection when there are multiple available output port selections.

D. RFA FAULT-TOLERANT ROUTING ALGORITHM

The purpose of the RFA fault-tolerant routing algorithm is to route packets along the minimum congestion path while bypassing the marked fault area. If there is no minimum path, select the direction with lighter fault. For any given 2D-mesh wired/wireless hybrid architecture, when (xd,yd) = (xc,yc), this indicates that the packet has reached destination node and can be forwarded to the node's PE core through local port. If (xd,yd) ≠ (xc,yc), the current node forwards the packet to its neighbors through the N/E/S/W/WI port. The RFA fault-tolerant routing algorithm picks out an optimal direction to transmit packets. Adaptive routing generally offers one to two ports to provide the shortest path routing. In order to avoid encountering a failure, we make following port selection rules: Rule 1. When the current node is on the same line as the destination node, and there is no severe fault in this direction, the minimum route is followed; when there is a severe fault in this direction, rule 2 is used; Rule 2. When there are two available port directions (eq1: if the current node is not on the same line as the destination node), the route with a weaker fault degree is priorly routed (Direction with serious fault is forbidden); Rule 3. When there are severe faults in both alternative directions, select the direction remaining in the priority area in Fig.3 (using a non-minimum routing path to route the packet. Note: This routing policy does not have route backtracking, because when all three directions are severe faults, the data packet is not allowed to pass to this node.). Rule 4: In the case of the same degree of failure, select the direction with lower congestion (since congestion usually occurs around faulty node, and congestion will be lower when the direction of the fault is lighter).

When there are two or more ports with the same candidate state, use the port priority selection mentioned above to select the direction with larger difference value of coordinate to ensure that the routing algorithm has the greatest degree of freedom. Fig.9 is a flow chart of the RFA routing algorithm.

V. HARDWARE IMPLEMENTATION

Fig.10 shows a typical 5-stage pipeline router architecture. The five phases are: buffer write (BW), route calculation (RC), virtual channel assignment (VA), switch assignment (SA), and link traversal (LT). These functions are divided into four phases: (BW/RC), (VA/SA), (ST), (LT). The fault management unit contains a link fault register for each port and is responsible for propagating (receiving) a fault status signal to the neighbor router (fault aware information). The RC logic uses this information to exclude possible output port candidates after incurring failure. The congestion management unit stores congestion information of the downstream router and is used to calculate the network state level relative to the communicated service. We consider buffer occupancy rate as primary indicator to measure congestion level and a secondary metric to sort all possible candidate output port RC.

The structure of the RFA is depicted in Fig.11, includes a preselected port module operating according to the current and destination node location, and a comparator component. The pre-selected port comparator processes the fault/congestion information in the network to accommodate a selection port for the optimal path.

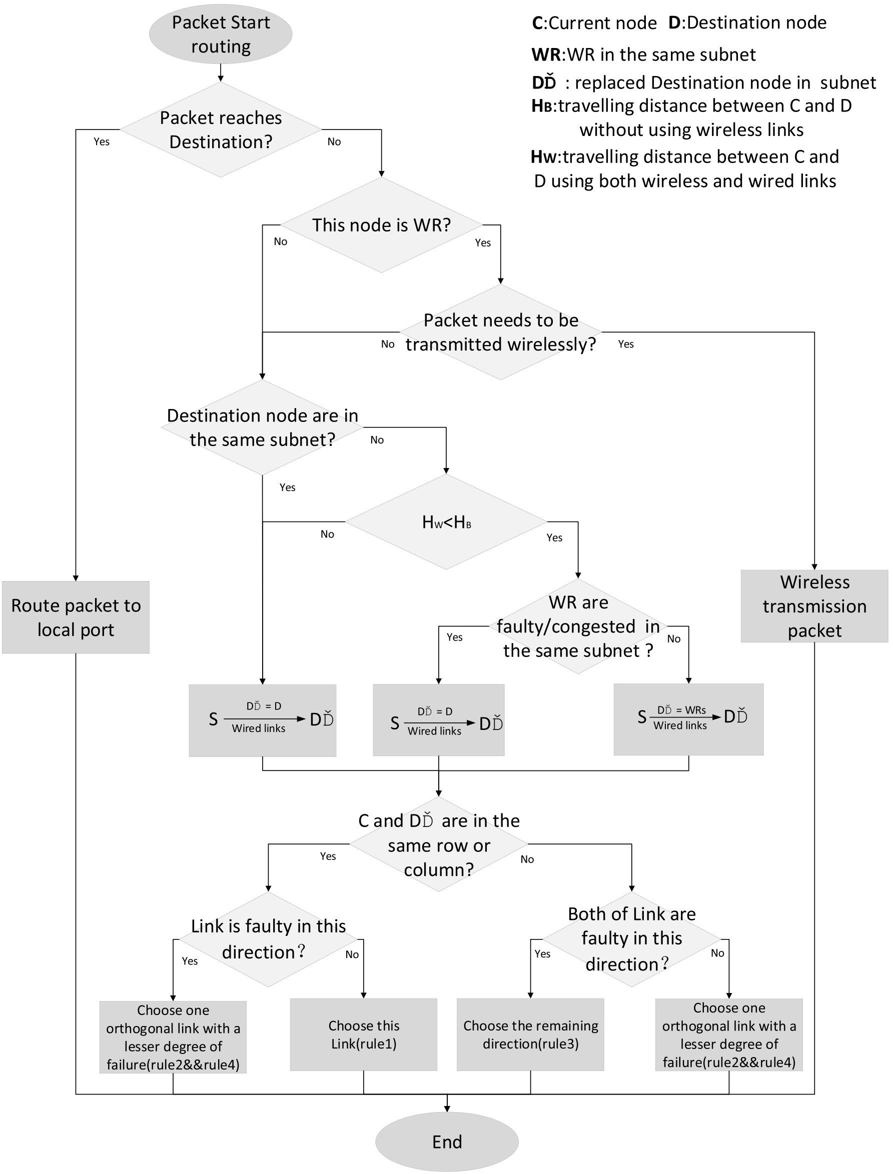

C:Current node D:Destination node
WR:WR in the same subnet
DĎ : replaced Destination node in subnet
HB:travelling distance between C and D without using wireless links
Hw:travelling distance between C and D using both wireless and wired links
Packet Start routing
Packet reaches Destination?
Yes
No
This node is WR?
No
Yes
Packet needs to be transmitted wirelessly?
No
Yes
Destination node are in the same subnet?
No
Yes
Hw<HB
No
Yes
WR are faulty/congested in the same subnet ?
Yes
No
Route packet to local port
Wireless transmission packet
S DĎ = D Wired links DĎ
S DĎ = D Wired links DĎ
S DĎ = WRs Wired links DĎ
C and DĎ are in the same row or column?
Yes
No
Link is faulty in this direction？
Yes
No
Both of Link are faulty in this direction？
Yes
No
Choose one orthogonal link with a lesser degree of failure(rule2&&rule4)
Choose this Link(rule1)
Choose the remaining direction(rule3)
Choose one orthogonal link with a lesser degree of failure(rule2&&rule4)
End

FIGURE 9. RFA routing algorithm flow chart.

VI. EXPERIMENTAL EVALUATION

Our experiments mainly evaluate the network performance, area overhead, and power consumption of RFA algorithm. We take three algorithms as counterparts in this paper. Scheme 1 is the traditional wired NoC; scheme 2 uses the FADyAD scheme proposed in [31]; scheme 3 uses HLAFT fault-tolerant routing [32]; scheme 4 is the two-hop-aware FoN routing algorithm [33]; Scheme 5 is the program of this paper (RFA).

A. PERFORMANCE ANALYSIS

Network performance simulation is performed by the extended on-chip network simulation tool Noxim [34]. All simulations are performed for 10,000 cycles, and 1000 clock cycles are used for system warm-up to achieve a relatively stable state. The basic parameters of the experiment are set as shown in Table 1. We set Random mode, Transpose mode, Shuffle mode and MMS mode to analyze the average delay and throughput of the network.

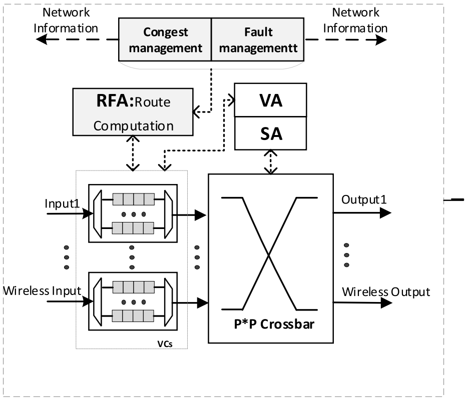


FIGURE 10. Router architecture of RFA.

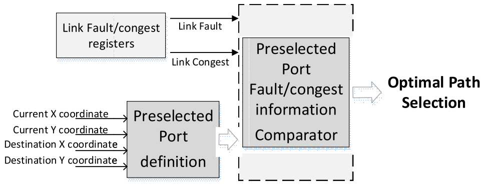


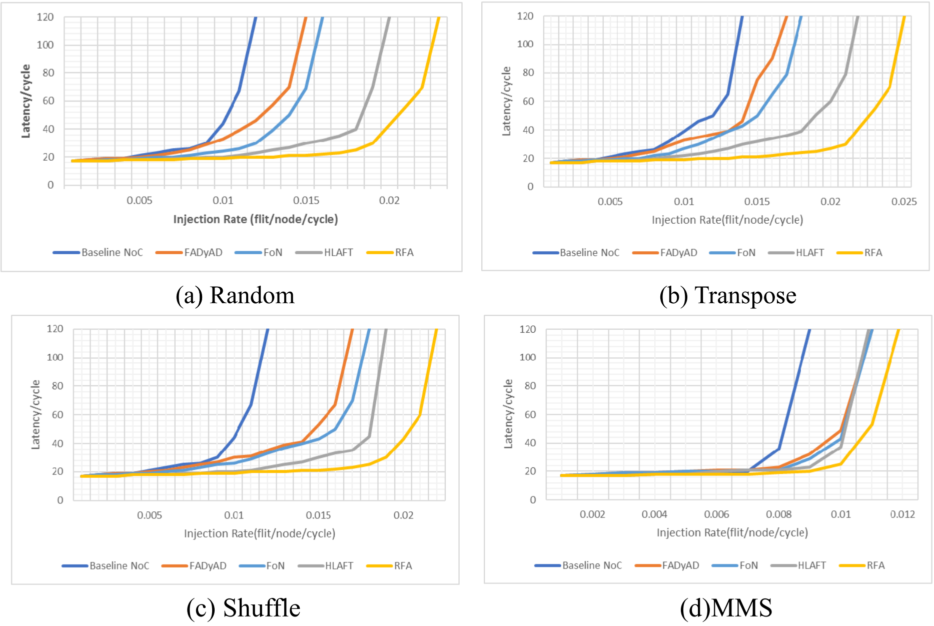
FIGURE 11. RFA structure.

TABLE 1. Basic parameters for experiment.

Parameter	Value
Topology	100 cores,4 subnets
Routing strategy	FADyAD, HLAFT, FoN, RFA
MAC	Token-Passed
Buffer Depth of router	3 and per 6 flit depth
WI buffer size	16 flit
Packet length	2-8 flit
Flit length	32 bit
Clock frequency	1 GHz
Wireless link bandwidth	32 Gbps
Traffic mode	Random、Transpose、Shuffle and MMS.

Latency and throughput are important indicators of network performance in NoC. The average delay is the average time data packet spend to traverse from source node to destination node. Network throughput is defined as the average number of slices of data received by each node in a single clock cycle. In the case of a 10*10 2D-mesh topology WiNoC, the performance of the analog RFA is compared without a failure rate. In Fig.12, when injection rate is low, delays of all schemes are similar, since packets in the network are not congested yet. As the injection rate increases, the probability of congestion of wireless nodes increases, and the delay of various schemes increases. It can be seen that the routing algorithm with congestion awareness has lower latency than other non-congestion perceptions. This is because in the on-chip network, the data traffic is unbalanced spatially and temporally. The node can sense surrounding load, and then perform routing to balance traffic distribution of the network. HLAFT will select ports with more path diversity for routing. However, if data packets of this node are selected in this way, the load on the port will be increased, and the congestion of the port will not be alleviated. Therefore, the congestion-aware routing algorithm is very suitable. Fig.13 shows the comparison of the throughput of the experimental routing algorithm. For the RFA algorithm, since the WR congestion/failure information is perceived by two hops, when the WR is congested, the wireless packets are not gathered around the WR. In two hops away, we have chosen a different path routing, balanced distribution of traffic around the WR, improving overall throughput of the network.

We put injection rate on the edge of saturation point (close but hasn't reach) and set link failure rate as 5%, 10%, and 20% respectively. The benefit of using this injection rate as a baseline value is that for non-faulty NoC systems (i.e., unsaturated), the latency and throughput of all routing algorithms vary significantly; this allows for fair performance evaluation of system performance over various failure link percentages. Fig.14 shows a comparison of experimental protocols in different flow modes. The RFA routing algorithm has lower average latency and higher throughput than other algorithms. Although FoN routing algorithm is a two-hop-aware routing algorithm, in the case of a high failure rate, the degree of failure beyond two hops is not perceived, and packet backtracking is still prone to occur, resulting in an increase in delay. When HLAFT selects a port with path diversity, its



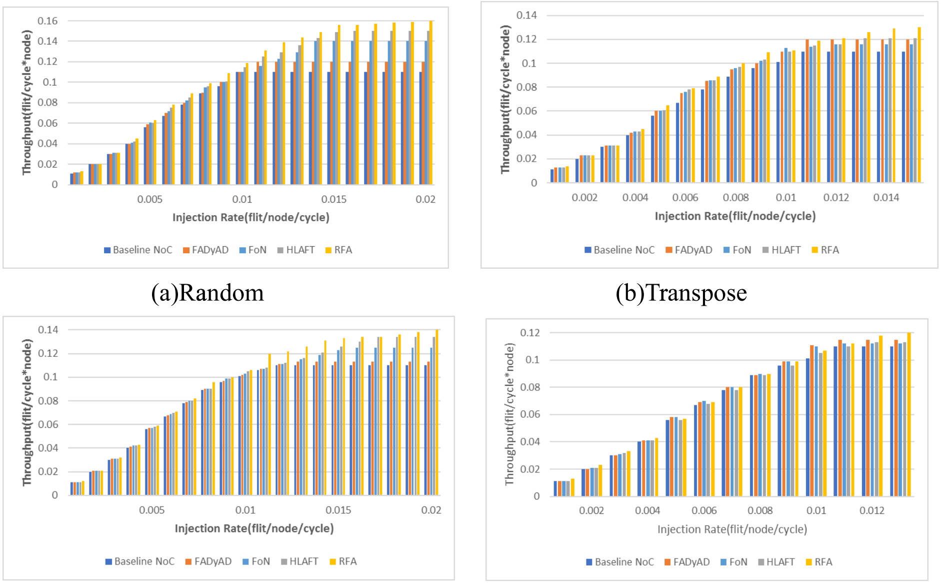
(a) Random

(b) Transpose

(c) Shuffle

(d)MMS

FIGURE 12. Average latency in no failure situation with different flow modes.



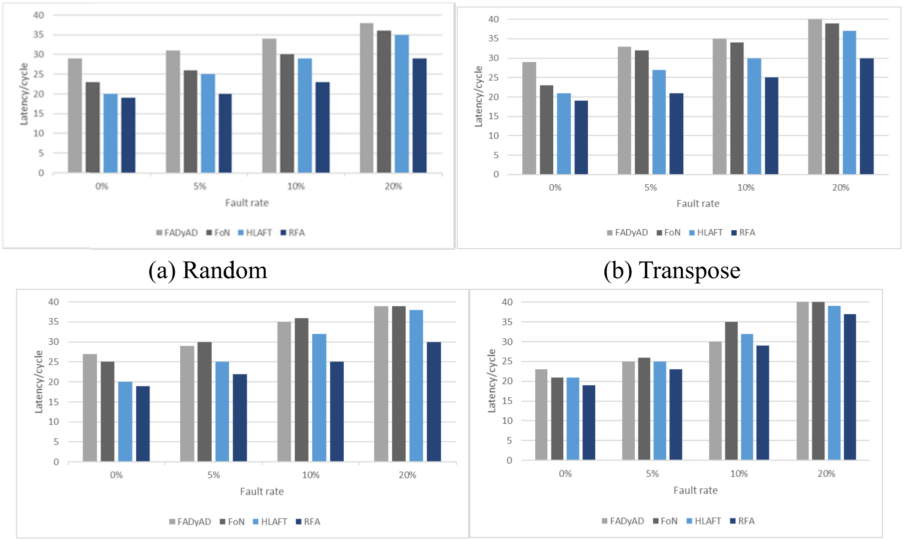
(a)Random

(b)Transpose

(c) Shuffle

(d)MMS

FIGURE 13. No-failure throughput in different traffic modes.



(a) Random

(b) Transpose

(c) Shuffle

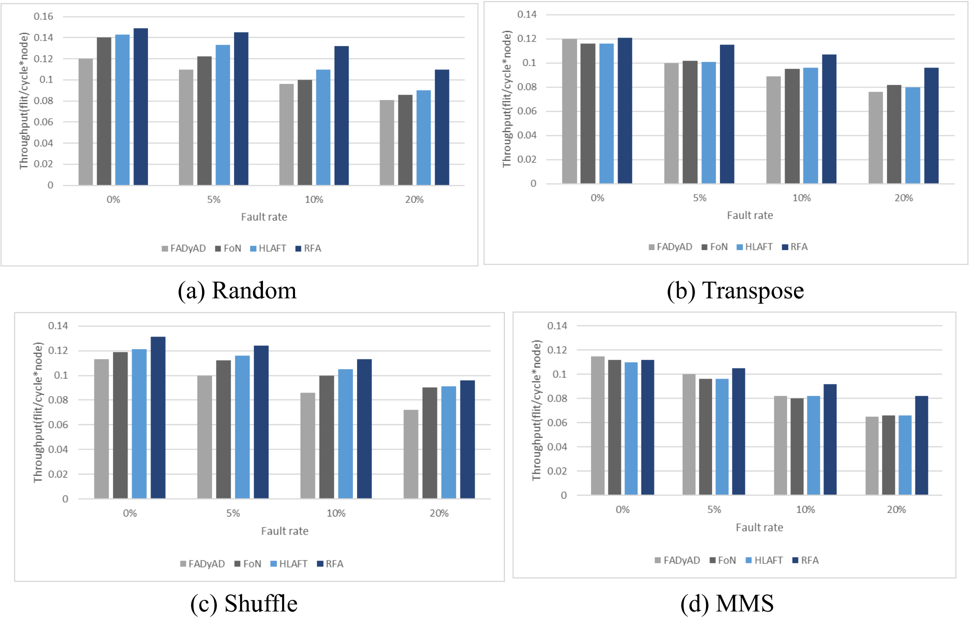
(d)MMS

FIGURE 14. Average delay in different traffic modes at 5%, 10%, 20% failure rate.

congestion and fault degree are not considered. The HLAFT routing algorithm is prone to routing to the faulty area, resulting in non-minimum Path routing, increasing latency. The RFA routing algorithm can detect congestion and fault areas. By combining the congestion and fault levels around the current node, the RFA routing algorithm selects an optimal path to reach the destination node. Therefore, this routing algorithm performs better when failure occurs.

Fig.15 shows the throughput comparison for each routing algorithm. When system has a faulty link, the throughput performance of all routing algorithm is reduced since the failed link no longer sends packets. As the failure rate increases, the rate of FADyAD, FoN, and HLAFT drops rapidly. In those routing algorithms packets are prone to be routed to the severely faulty area, causing blockage of nodes in the network and reducing throughput. The RFA routing algorithm selects the direction routing packet with a small probability of failure area, so that the data packet can reach the destination node more quickly, reducing the rate of throughput drop due to the increase of the failure rate, and the system can maintain in a stable state.

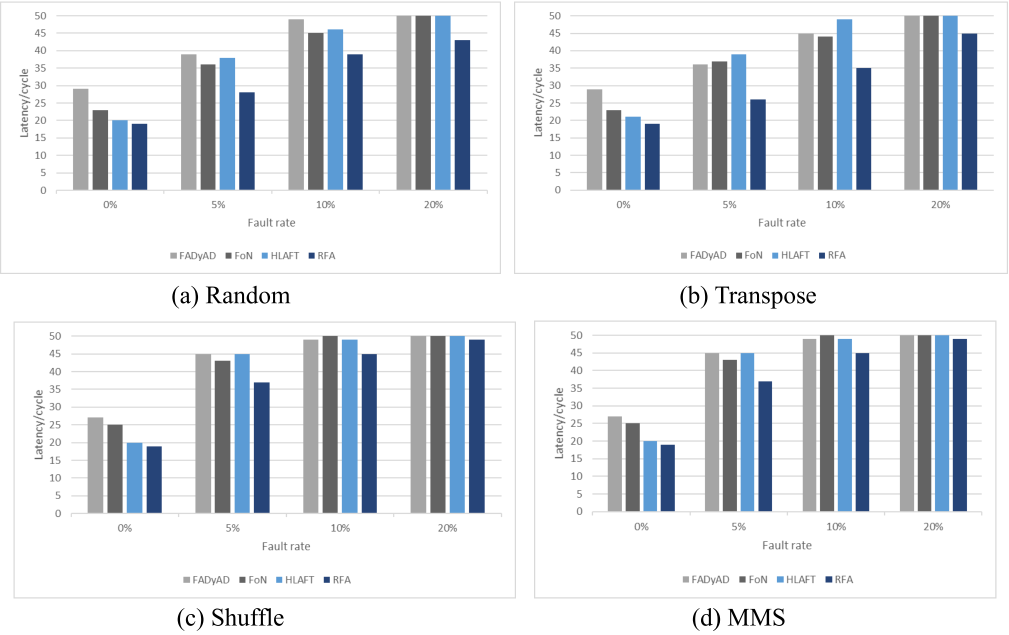
We explore the impact on system performance with each algorithm under 0%, 5%, 10%, 20% failure rate, as shown in figure.16 and figure.17. Wireless nodes function as communication center so when they incur failures whole system performance degrades severally. In this paper we focus on



(a) Random (b) Transpose

(c) Shuffle (d) MMS

FIGURE 15. **Throughput at 5% failure rate, different traffic patterns.**

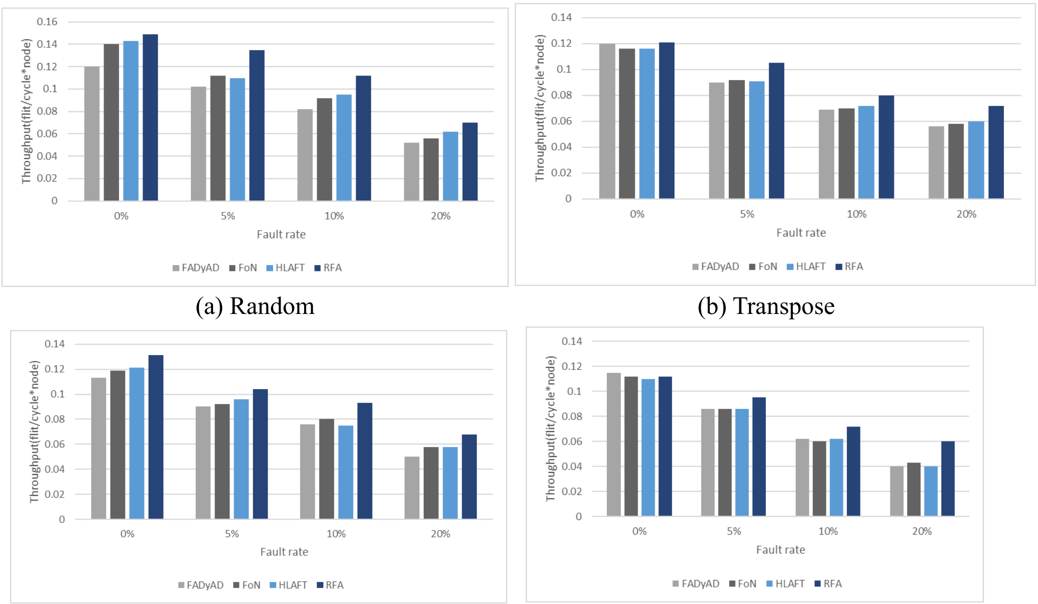


(a) Random (b) Transpose

(c) Shuffle (d) MMS

FIGURE 16. **Average delay in a fault WI and 5%, 10%, 20% failure rate, different flow patterns.**

communication level so processing delay in PE cores is not considered. In Fig.16, it can be seen that when the wireless node fails, the delay in the network increases very fast, because a large number of data packets arrive at the wireless node, and cannot be transmitted by wireless channels, thereby generating congestion for the wireless node, causing overall load uneven. The general routing algorithm only perceives the wireless node congestion by one hop, but since wireless channel is usually shortcut and thus attracts most packets gathering, routers around wireless node will get congested. The RFA algorithm senses the congestion/failure situation of the wireless node in two hops. When the data packet is routed to the node within its two hops, other effective path routes can be selected, which reduces the load pressure of the wireless node and its surroundings and balance the overall load largely. Fig.17 shows the change in throughput at different failure rates. Since the RFA senses the fault information of the two nodes of the wireless node, the wireless nodes in the network are not congested, maintaining robustness of the system and load balance of the network.



(a) Random

(b) Transpose

(c) Shuffle

(d)MMS

FIGURE 17. Throughput in a fault WI and 5%, 10%, 20% failure rate, different traffic patterns.

TABLE 2. Router area overhead and power consumption.

Routing Algorithm	Area overhead(um^2)			power consumption(mW)
	Baseline Router	Additional area	Additional area (WR)	
FADyAD	49987	2001	×	×
FoN	49987	5069	×	23.34
HLAFT	50519	6001	×	21.12
RFA	51236	4035	1036	22.56

B. AREA AND POWER CONSUMPTION

To evaluate area and power cost, we use Synopsys' Design Compiler with Verilog HDL to build hardware simulation. In terms of wireless communication, the Zig-Zag antenna used in this paper has an area of approximately 0.19 and a millimeter-wave transceiver area of approximately 0.16 [27].

Table 2 shows the area and power consumption of the four schemes under the 65nm standard library. It can be seen from the table that the area overheads of the four schemes are relatively close. For increased area overhead, since FADyAD, FoN, and HLAFT all need to increase signal lines and registers to realize fault perception, FoN and HLAFT routers need two-hop sensing, so the additional area is larger. In this paper, the wired router adopts one-hop sensing, and the wireless router adopts the two-hop sensing method to obtain the trade-off of area and performance. When the fault is perceived, the transmitted signal line and the received register are twice as large as other schemes. The overall area of the solution in this paper has increased by 9.89%. Compared with other solutions and the performance improvement, it's an acceptable area cost. In terms of power consumption, both fault area awareness and congestion control require a control management and arbitration unit. The power consumption of this solution is basically the same as other solutions.

VII. CONCLUSION

As a promising interconnection method in on-chip multi-core systems, WiNoC improves network performance of traditional wired NoC. In this paper, a novel RFA routing algorithm is proposed for link congestion/failure problem in WiNoC. Based on reginal fault awareness, the router can understand the surrounding fault status. Congestion control is fused to the RFA routing algorithm to alleviate congestion around failure. We trade off performance and overhead by adopting 2-hop awareness for wireless nodes and 1-hop awareness for wired nodes. And the node status is not defined merely based on its own faulty condition but with a "double sensing" mechanism. Through regional fault status and congestion information, optimal routing path can be selected for the packet. Combining one-hop awareness of wired routers with the two-hop perception of wireless routers yields a good trade-off between performance and overhead. The experimental results show that the proposed scheme greatly improves the robustness of the system compared to the comparison object within the acceptable range of area and power consumption.

REFERENCES

[1] L. Benini, G. De Micheli "Networks on chip: A new paradigm for systems on chip design," in *Proc. Automat. Test Eur. Conf. Exhib.*, 2002, pp. 418–419.

[2] M. M. Shokoofeh and J. J. M. Ali, “A load-balanced congestion-aware routing algorithm based on time interval in wireless network-on-chip,” *J. Ambient Intell. and Humanized Comput.*, vol. 10, pp. 1–14, Sep. 2018.

[3] *International Technology Roadmap for Semiconductors (ITRS)*, Semicond. Ind. Assoc., Wilson, Linda, 2013.

[4] C. H. Chao, K. Y. Jheng, H. Y. Wang, J.-C. Wu, and A.-Y. Wu “Traffic- and thermal-aware run-time thermal management scheme for 3D NoC systems,” in *Proc. 4th ACM/IEEE Int. Symp. Netw. Chip*, May 2010, pp. 223–230.

[5] A. Joshi, C. Batten, Y.-J. Kwon, S. Beamer, I. Shamim, K. Asanovic, and V. Stojanovic “Silicon-photonic clos networks for global on-chip communication,” in *Proc. 3rd ACM/IEEE Int. Symp. Netw. Chip*, May 2009, pp. 124–133.

[6] M. F. Chang, J. Cong, A. Kaplan, M. Naik, G. Reinman, E. Socher, and S.-W. Tam, “CMP network-on-chip overlaid with multiband RF-interconnect,” in *Proc. IEEE 14th Int. Symp. High Perform. Comput. Archit.*, Feb. 2008, pp. 191–202.

[7] 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 Trans. Parallel Distrib. Syst.*, vol. 26, no. 12, pp. 3289–3302, Dec. 2015.

[8] D. W. Matolak, S. Kaya, and A. Kodi “Channel modeling for wireless networks-on-chips,” *IEEE Commun. Mag.*, vol. 51, no. 6, pp. 180–186, Jun. 2013.

[9] W.-H. Hu, C. Wang, and N. Bagherzadeh, “Design and analysis of a mesh-based wireless network-onchip,” *J. Supercomput.*, vol. 71, pp. 483–490, Feb. 2012.

[10] M. RadetzkiM, C. Feng, X. Zhao, and A. Jantsch, “Methods for fault tolerance in networks-on-chip,” *ACM Comput. Surv.*, vol. 46, pp. 1–38, Oct. 2013.

[11] P. Wettin, A. Vidapalapati, A. Gangul, and P. P. Pande, “Complex network-enabled robust wireless network-on-chip architectures,” *ACM J. Emerg. Technol. Comput. Syst.* vol. 9, pp. 1–19, Sep. 2013.

[12] J. Zhou, H. Li, T. Wang, S. Li, Y. Wang, and X. Li, “TWiN: A turn-guided reliable routing scheme for wireless 3D NoCs,” in *Proc. IEEE 24th Asian Test Symp. (ATS)*, Nov. 2015, pp. 85–90.

[13] B. Bhowmik, J. K. Deka, S. Biswas, and B. B. Bhattacharya, “A topology-agnostic test model for link shorts in on-chip networks,” in *Proc. IEEE Int. Conf. Syst., Man, (SMC)*, Oct. 2016, pp. 4561–4566.

[14] K. Chang, S. Deb, A. Ganguly, X. Yu, S. P. Sah, P. P. Pande, B. Belzer, and D. Heo, “Performance evaluation and design trade-offs for wireless network-on-chip architectures,” *ACM J. Emerg. Technol. Comput. Syst.*, vol. 8, no. 3, pp. 1–25, 2012.

[15] P. Wettin, P. P. Pande, D. Heo, B. Belzer, S. Deb, and A. Ganguly, “Design space exploration for reliable mm-wave wireless NoC architectures,” in *Proc. IEEE 24th Int. Conf. Appl.-Specific Syst., Archit. Processors*, Jun. 2013, pp. 79–82.

[16] A. Rezaei, M. Daneshtalab, D. Zhao, F. Safaei, X. Wang, and M. Ebrahimi, “Dynamic application mapping algorithm for wireless network-on-chip,” in *Proc. 23rd Euromicro Int. Conf. Parallel, Distrib., Netw.-Based Process.*, Mar. 2015, pp. 421–424.

[17] A. Rezaei, M. Daneshtalab, M. Palesi, and D. Zhao, “Efficient Congestion-aware scheme for wireless on-chip networks,” in *Proc. 24th Euromicro Int. Conf. Parallel, Distrib., Netw.-Based Process. (PDP)*, Feb. 2016, pp. 742–749.

[18] A. Ganguly, P. Pande, B. Belzer, and A. Nojeh, “A unified error control coding scheme to enhance the reliability of a hybrid wireless network-on-chip,” in *Proc. IEEE Int. Symp. Defect Fault Tolerance VLSI Nanotechnol. Syst.*, Oct. 2011, pp. 277–285.

[19] V. Vijayakumaran, M. P. Yuvaraj, and N. Mansoor, “CDMA enabled wireless network-on-chip,” *ACM J. Emerg. Technol. Comput. Syst.*, vol. 10, no. 4, pp. 1–20, 2014.

[20] A. Dehghani and K. Jamshidi, *A Fault-Tolerant Hierarchical Hybrid Mesh-Based Wireless Network-on-Chip Architecture for Multicore Platforms*. Norwell, MA, USA: Kluwer, 2015.

[21] R. Ausavarungnirun, K. K.-W. Chang, C. Fallin, and O. Mutlu, “Adaptive cluster throttling: Improving high-load performance in bufferless on-chip networks,” Comput. Archit. Lab, Carnegie Mellon Univ., Pittsburgh, PA, USA, SAFARI Tech. Rep. 6, 2011.

[22] G. P. Nychis, C. Fallin, T. Moscibroda, O. Mutlu, and S. Seshan “On-chip networks from a networking perspective: Congestion and scalability in many-core interconnects,” in *Proc. ACM Conf. Appl., Technol., Archit., Protocols Comput. Commun. (SIGCOMM)*. New York, NY, USA: ACM, 2012, pp. 407–418.

[23] Y. Ouyang, J. Yang, and K. Xing, “An improved communication scheme for non-HOL-blocking wireless NoC,” *Integration*, vol. 60, pp. 240–247, Jan. 2018.

[24] 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 Trans. Comput.*, vol. 57, no. 9, pp. 1230–1245, Sep. 2008.

[25] S. Deb, K. Chang, X. Yu, S. P. Sah, M. Cosic, A. Ganguly, P. P. Pande, B. Belzer, and D. Heo, “Design of an Energy-Efficient CMOS-Compatible NoC Architecture with Millimeter-Wave Wireless Interconnects,” *IEEE Trans. Comput.*, vol. 62, no. 12, pp. 2382–2396, Dec. 2013.

[26] S. Deb, A. Ganguly, P. P. Pande, B. Belzer, and D. Heo “Wireless NoC as interconnection backbone for multicore chips: Promises and challenges,” *IEEE J. Emerg. Sel. Topics Circuits Syst.* , vol. 2, no. 2, pp. 228–239, Jun. 2012.

[27] X. Yu, J. Baylon, P. Wettin, D. Heo, P. P. Pande, and S. Mirabbasi, “Architecture and design of multichannel millimeter-wave wireless NoC,” *IEEE Des. Test*, vol. 31, no. 6, pp. 19–28, Dec. 2014.

[28] A. Ghiribaldi, D. Ludovici, and F. A. TriviO, “Complete self-testing and self-configuring NoC infrastructure for cost-effective MPSoCs,” *ACM Trans. Embedded Comput. Syst.*, vol. 12, no. 4, p. 1, 2013.

[29] Y. Ouyang, J. Da, and X. Wang, “A TSV fault-tolerant scheme based on failure classification in 3D-NoC,” *J. Circuits, Syst., Comput.*, vol. 26, no. 04, p. 19, 2017.

[30] C. Feng, Z. Lu, A. Jantsch, M. Zhang, Z. Xing, “Addressing transient and permanent faults in NoC with efficient fault-tolerant deflection router,” *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 21, no. 6, pp. 1053–1066, Jun. 2013.

[31] A. Mehranzadeh, A. Khademzadeh, and A. Mehran, “Fadyad- fault and congestion aware routing algorithm based on dyad algorithm,” in *Proc. 5th Int. Symp. Telecommun.*, Dec. 2010, pp. 274–279.

[32] A. Ben Ahmed and A. Ben Abdallah, “Graceful deadlock-free fault-tolerant routing algorithm for 3d network-on-chip architectures,” *J. Parallel Distrib. Comput.*, vol. 74, no. 4, pp. 2229–2240, 2014.

[33] C. Feng, Z. Lu, A. Jantsch, J. Li, and M. Zhang, “FoN: Fault-on-neighbor aware routing algorithm for networks-on-chip,” in *Proc. 23rd IEEE Int. SOC Conf.*, Sep. 2010, pp. 441–446.

[34] V. Catania, A. Mineo, S. Monteleone, M. Palesi, and D. Patti, “Noxim: An open, extensible and cycle-accurate network on chip simulator,” in *Proc. IEEE 26th Int. Conf. Appl.-Specific Syst., Archit. Processors (ASAP)*, Jul. 2015, pp. 162–163.

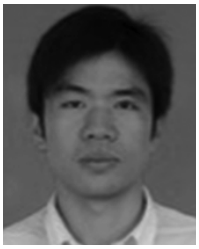
[35] R. Akbar and F. Safaei, “A novel congestion-aware and adaptive routing algorithm in mesh-based networks-on-chip with segmentation,” in *Proc. 19th Int. Symp. Comput. Archit. Digit. Syst. (CADS)*, Dec. 2017, pp. 1–6.

YIMING OUYANG received the bachelor’s, master’s, and Ph.D. degrees from the Hefei University of Technology, in 1984, 1991, and 2013, respectively. He is currently a Professor with the Hefei University of Technology, and a Leading Expert in research. His main research directions are NoC, on-chip systems (SoC), integration and testing of embedded systems, and automation of digital system design.

QI WANG graduated from Jilin University, in 2016. He received the bachelor’s degree from the Stevens Institute of Technology, in 2018. He is currently pursuing the Ph.D. degree with the Hefei University of Technology. He was an Exchange Student majoring in computer and system science at Stockholm University, in 2014. His main research fields are network-on-chip, machine learning, fault tolerance, and NoC-based neural network accelerators.

MENGXUAN RU received the B.Sc. degree from the Hefei University of Technology, in 2016, where he is currently pursuing the M.Sc. degree. His main research interests include network-on-chip and wireless network-on-chip.

HUAGUO LIANG received the bachelor's and master's degrees from the Hefei University of Technology, in 1982 and 1989, respectively, and the Ph.D. degree from the University of Stuttgart, in 2003. His main research fields are SoC and VLSI design and test.



JIANHUA LI received the B.S. degree from the Department of Computer Science and Technology, Anqing Normal University, Anhui, China, in 2007, and the Ph.D. degree from the School of Computer Science and Technology, University of Science and Technology of China, in 2013. He is currently a full-time Lecturer with the School of Computer and Information, Hefei University of Technology, Anhui. His research interests include multicore memory systems, emerging nonvolatile memories, and on-chip networks.