Indonesian Journal of Electrical Engineering and Computer Science
Vol. 27, No. 2, August 2022, pp. 1051~1061
ISSN: 2502-4752, DOI: 10.11591/ijeecs.v27.i2.pp1051-1061

Thermal-aware directional and adaptive routing algorithm for 3D network-on-chip

Muhammad Kaleem[1,2], Ismail Fauzi Isnin[1]
[1]School of Computing, Faculty of Engineering, Universiti Teknologi Malaysia, Johor Bahru, Malaysia
[2]Department of Computer Science and Information Technology, University of Sargodha, Sargodha, Pakistan

Article Info

Article history:

Received Sep 11, 2021
Revised May 27, 2022
Accepted Jun 11, 2022

Keywords:

Network traffic
Routing algorithms
System-on-chip
Thermal-aware
Very large scale integrations

ABSTRACT

Due to the tier architecture of 3D network-on-chip (3D-NoC), reducing the thermal hotspot within the chip is challenging as a cooling mechanism that lies merely on the single side of a chip. High power density in 3D NoC is responsible for reliability degradation and thermal difficulties. Thermal-aware routing becomes substantial to handle thermal difficulties and diffusion of heat to the cooler regions. Thermal-aware routing focuses on bypassing hotspot areas by selecting cooler areas. Existing thermal-aware routing algorithms adopt slightly cooler but longer and extended paths, due to lack of ability to know the proximity of the destination's location, which aggravate thermal issues. This work presents a novel thermal-aware directional and adaptive routing algorithm. Objective of the proposed algorithm is to strive to find the best possible neighbour to reach closer to the proximity of the destination. The proposed algorithm can adaptively choose any suitable neighbour that can lead packets closer to the destination at each intermediate node. The performance of the proposed algorithm is evaluated and compared with existing thermal-aware routing algorithm in a simulator environment. Simulation results demonstrate that the proposed method outperformed its counterpart in terms of average delay with 11-26% improvement, total hop counts with 8-24% reduction under various traffic conditions and improvement in overall thermal profiling of the chip.

Corresponding Author:

Muhammad Kaleem
School of Computing, Faculty of Engineering, Universiti Teknologi Malaysia
Johor Bahru, Malaysia
Email: kaleem.muhammad@graduate.utm.my

1. INTRODUCTION

Stacking of dies in three-dimensional integrated circuit (3D-IC) has reduced interconnection delays. 3D network-on-chip (3D NoC) based designs are expected to offer improved performance while using fewer data transfer connections and consuming less power [1], [2]. In conventional NoC architecture, each tile comprises a network interface, processing element (PE), and a router [3], [4]. A high switching activity in routers is responsible for thermal hotspots in NoC [5]. Chip cooling mechanism, also commonly known as a heat sink, is utilized merely on a single lateral of the multiple layers in 3D NoC; therefore, layers that lie subsequently away from the cooling mechanism have longer heat dissipation paths. Hence, the possibility of the thermal hotspots increases especially in layers subsequently away from the sink [6]. Thermal hotspots exacerbate the failure mechanism in 3D NoC [7], putting an additional burden on chip cooling costs and system reliability. To eliminate thermal hotspots while retaining performance, designers must understand the influence of temperature changes on the systems [8]. Due to increased power density, thermal issues are a growing concern in modern microelectronics. Thermal issues make systems more vulnerable to temperature

effects such as reliability, leakage and delay [9]. To balance temperature distribution various methods have been proposed such as, floor-plan optimization [10], [11], thermal-aware application mapping [12], and thermal-aware routing [13], [14]. There are two types of thermal-aware routing algorithms, temporal thermal-aware routing algorithms and spatial thermal-aware routing algorithms. Temporal thermal-aware routing algorithms can reduce temperatures on-chip by dynamically adjusting frequencies, voltages, or clock cycles, but temporal thermal-aware routing algorithms reduce overall system performance. Spatial thermal-aware routing algorithms are more promising as they are intended to reduce thermal hotspot by distributing traffic away from heated areas [15]. There are further two types of spatial thermal-aware routing algorithms: reactive thermal-aware routing algorithms and proactive thermal-aware routing algorithms [16].

In case of throttling situations, reactive thermal-aware routing algorithms provide alternative cool paths to reach its destination. As far as proactive thermal-aware routing is concerned, proactive thermal-aware routing algorithms prevent nodes from getting throttled by managing NoC traffic prior to emergency situations. Proactive thermal-aware routing can delay emergency states of nodes but cannot guarantee throttling freeness. Recently, thermal-aware routing algorithms have gained importance within the research community due to high switching activity in the routers. An efficient routing algorithm reduces traffic hotspot formation, congestion and packet delays. Each 3D-NoC routing algorithm focuses on deadlock freedom, link assignment for every packet and thermal management. GTDAR is a game theory based thermal delay-aware routing [17] to orchestrate traffic and thermal situations more accurately to restrict and convert the temperature issue into traffic issue, GTDAR transfers the long term thermal information into short term traffic information. GTDAR has an unbalanced traffic load in the network. Immediate neighbourhood temperature (INT) is an adaptive routing algorithm [18].

INT balances temperatures by routing packets along low temperature paths in 3D-NoC. INT exploits locally available and adjacent routers temperature information to select its output port for incoming packets. INT lacks a global view of temperature; it only takes decisions on the immediate node temperatures. Immediate node temperatures can mislead packets away from their destination. In location-based aging-resilient Xy-Yx (LAXY) [15] NoC routers are separated in two sub-groups. Packets are routed using YX routing in one group whereas, other group uses XY routing algorithm to transfer packets between source and destination. This configuration reduced load from the central nodes. LAXY only focuses on congestion in the center of the network. Each packet in adaptive thermal-aware routing (ATAR) [19] traverses the network based on the weighted cost model calculation for each packet. To choose the best neighbour for dispatching the packet, the cost for the potential next neighbour is calculated. Each neighbour will be visited only once and marked as visited. It keeps marking nodes as visited; it keeps on reducing its options to reach the destination. It has the destination address but it does not have the proximity of the location of the destination. Just bases on least cost, it looks around and has a little chance of reaching its destination directly even if NoC is thermally stable. Due to poor traffic distribution and a lack of cooling mechanism between the layers, it is challenging to deal with heat dissipation issues in 3D-NoC. Longitudinal exclusively adaptive or deterministic (LEAD) [20] distributes load equally across layers and nodes to avoid transistor biasing from overheating. To solve thermal issues, INT [18] just measures the temperature of the immediate neighbourhood. According to the literature, other aspects such as surrounding node temperatures, congestion detection, shortest path length, and the next router queue length must be considered in addition to thermally aware selection. ATAR [19] achieves adaptability by allowing choosing any of the neighbours. This adaptability without its direction vector, forces packets to choose unwanted nodes based on minimum cost. Choosing a random node leads it to move away from its destination. On some occasions it even passes by its destination and takes a longer route and then manages to reach its destination. Taking longer routes can raise issues like congestion, power leakage leads to thermal instability and temporary faults. Hence balance between temperature and path length is required.

Due to the poor traffic distribution and lack of heat sinks among the layers, it is difficult to handle the heat dissipation issue in 3D NoC. Many routing algorithms that claim to deal with thermal issues effectively. The center of the network is more likely to observe thermal issues, so one of the most common methods to address thermal issues is detouring network traffic away from the center. Detouring of traffic results in slightly cooler paths, but at the same time increases path length. Longer paths result in more hops and more intermediate traffic making its way to reach its destination. Higher intermediate traffic causes more congestion, induces thermal, and delay issues.

This paper proposes a new approach that takes into account congestion, temperature, and most crucially, allows packets to choose their next neighbour adaptively that will lead them closer to the proximity of the destination. The contributions of this work are highlighted: i) a novel proactive thermal-aware directional and adaptive routing (TADAR) is proposed. It is a technique that strives to find the best possible neighbour to reach near to the proximity of the destination, ii) at each intermediate node proposed technique looks for a suitable adaptive neighbour that can lead the packet one step closer to the destination, and iii)

proposed technique is highly adaptive in the beginning. It becomes deterministic just before reaching its destination. TADAR is simulated using the access Noxim simulator.

The rest of the paper is organized as follows. The detailed method of the proposed routing algorithm under various source-destination scenarios is expressed in section 2. The results of the simulations were presented in section 3. Finally, in section 4, this work has been concluded.

2. RESEARCH METHOD

The increased power demand in NoC causes the nodes to heat up. Even the tiniest of undesired processes can contribute to an increase in temperature, resulting in a crisis. As a result, excessive and uncontrolled network packet movement should be avoided. Consider the following situation: when the router is currently busy transmitting packets towards the required output port, incoming packets must wait for their turn in the input buffer of the packet queue. These extra packets stall in the input buffers, accelerating the thermal aggravation process. This paper presents an approach that takes temperature and congestion into account while also allowing packets to adaptively choose the next neighbour that will lead them closer to the destination.

2.1. Selection of candidate node

3D NoC is a three-dimensional network in x, y and z axis respectively. This layered architecture reduces distances between end-to-end nodes by providing alternative paths. Maximum neighbours of a node in two-dimensional (2D) networks are up to four. As far as 3D NoC is concerned, the maximum number of neighbouring nodes can be six, which brings more options to choose from while routing. The center node of the 3D NoC in Figure 1 has six neighbours in each direction i.e. south, west, east, south, up and down directions. Suppose the center node is an intermediate node responsible for forwarding packets to any of its neighbours based on the minimum temperature. As the destination is situated in the north-east of the intermediate node in the upper tier of Figure 1, hence it should check the temperatures of the relevant neighbours only. Relevant neighbours in this case are east, north and up. The center node will choose the node with the minimum temperature among the relevant neighbours and forward its packet to that neighbour. Similar operations will be performed at the next intermediate neighbour until it reaches its destination.

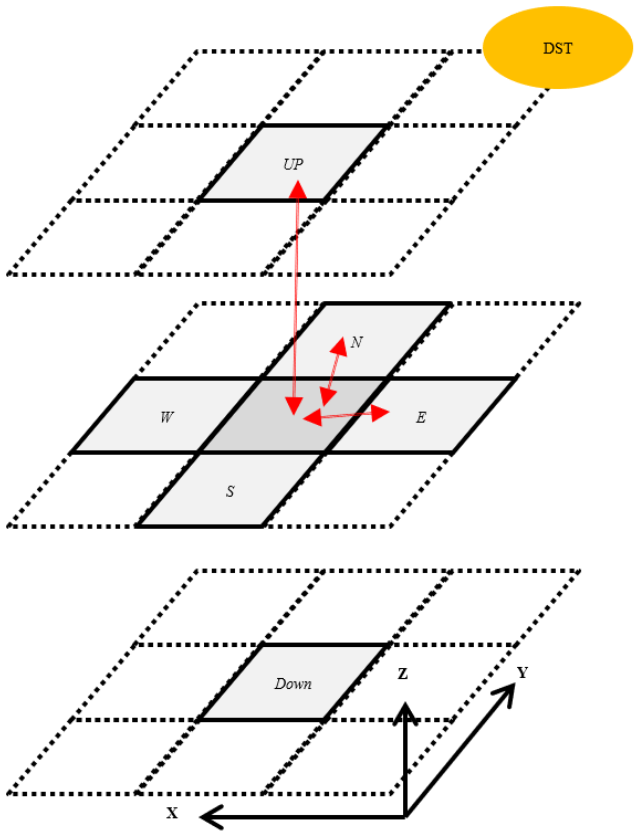

Figure 1. Central node neighbours of 3D NoC

Various scenarios of the proposed technique are depicted in Figure 2. Source and destination are in various rows and columns. Figure 2(a) shows that node 5 generates a packet for node 3 available in the north-west of the source node in the same layer. To reach node 3, the packet has more than one path option to reach closer to its destination. The next suitable candidate node can be found at each intermediate node without exceeding hops. In Figure 2(b), source node 3 is in the top row and destination node 9 is in last row. Again it can witness multiple available dimensions and paths to reach its destination. NoC has the potential to rapidly concentrate the center of the network increasing congestion. It can be observed that out of many paths there are some paths which are laying on the edge of the network. So, in case of heavy congestion it can still reach its destination using less congested paths. Figure 2(c) shows source is node 8 and its destination is node 6. The proposed algorithm has several possible ways to reach its destination, making it an adaptive routing algorithm. At the same time, it can also observe that packets do not take excessive path length by going away from the destination in any cases presented. In fact, it is trying to reach its destination by choosing the best possible low concentrated region.

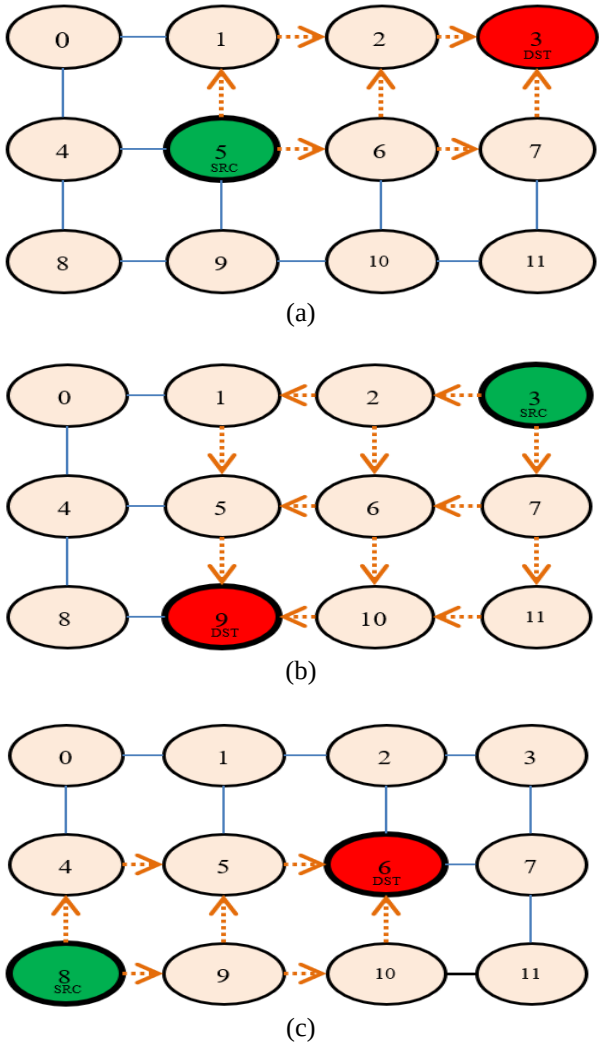

Figure 2. Various possible communication paths in the same layer (a) source in center row and destination on upper row, (b) source in top row and destination on bottom row, and (c) source in bottom row and destination in center row

The layered architecture in 3D NoC reduces distance between end to end nodes by providing alternative short paths. Consider the destination node is situated in a different layer, as shown in Figure 3. The source node is 21, whereas the destination node is 6 in the other tier. At node 21, immediate neighbours are 9, 17 and 22. All neighbours have the potential to carry the packet one step closer to the destination. Now

node 21 will decide the next intermediate node based in the least cost. If it chooses 22, it has neighbours of 23, 18, 10, and 21. Node 21 cannot be chosen because it was initiating node and node 23 will take the packet away from the destination; hence the options to choose from are node 18 or node 10 according to the proposed technique. Consider if it chooses node 18 based on least cost, it can be seen that from node 18, it has only one possible valid option to reach its destination. If the destination is one hop away from the intermediate node, then there is only one possible path available in the proposed technique. In the initial phase of the proposed technique, it is highly adaptive as it can choose any neighbouring node in any dimension but towards the destination in order to allow packets to move towards the destination. As it comes near to its destination it becomes highly directional. It is possible that this last hop may not be feasible in terms of cost but it eventually helps deliver one packet to its destination and reduces the number of transmitting packets over the network. Hence it plays an important role in reducing the congestion, especially from the center of the network.

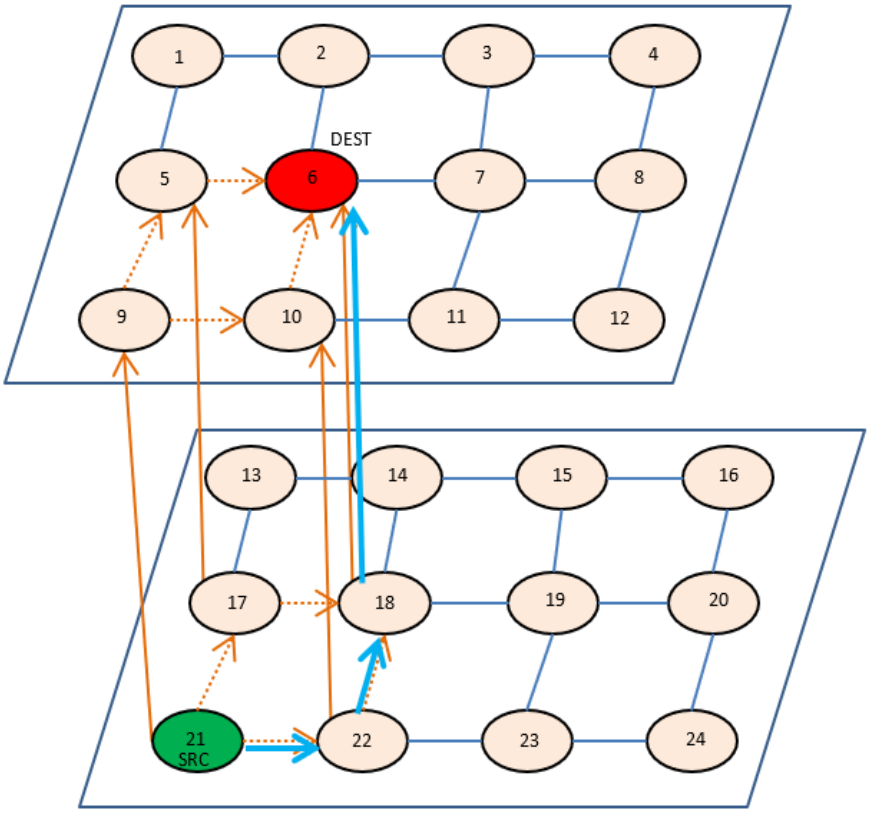

Figure 3. An example of destination in another layer

2.2. Thermal-aware directional and adaptive routing

TADAR algorithm takes the source node donated by s_node, the destination node donated by d_node, and the route data, presented in Algorithm 1. Route data include node parameters such as L, T, Q, W. The method will first examine the addresses of the source and destination nodes. If the source node and destination node are the same then, it will be terminated by returning direction Local in [line 3-4]. It can choose any neighbouring node in any dimension but has the potential to take the packet closer to its intended destination. As it reaches closer to the destination, the valid neighbour reduces to direct packets to reach destination [line 8-19]. Function getNeighbourID() returns neighbour ID in the requested direction. In [line 20-23] the corresponding edges are obtained for each available direction. The weighted cost sum is then calculated. If the calculated cost is less than current cost, the cost matrix is updated in [line 24-29], the cost model considered in this work is similar to [19] in [line 26]. New intermediate node donated by i_node, will be determined based on minimum cost and direction of minimum cost node is pushed in directions.

```
Algorithm 1. TADAR
function TADAR( s_node , d_node, route_data)
set opt_Nei[6] ← -1
 if s_node= d_node then
 directions ← direction_local
 else
    set i_node ← s_node
    while i_node ≠ d_node  do
        if d_node.x > i_node.x then
            set opt_Nei[1] ← getNeighbourID (i_node, EAST)
       else if d_node.x < i_node.x then
           set  opt_Nei[3] ← getNeighbourID (i_node, WEST)
       if d_node.y < i_node.y then
           set opt_Nei[0] ← getNeighbourID (i_node, NORTH)
       else if d_node.y > i_node.y then
           set opt_Nei[2] ← getNeighbourID (i_node, SOUTH)
       if d_node.z > i_node.z then
           set opt_Nei[5] ← getNeighbourID (i_node, DOWN)
       else if d_node.z > i_node.z then
           set opt_Nei[4] ← getNeighbourID (i_node, UP)
       For k ϵ valid_directions do
           set candidate_idn ← opt_Nei[k]
           if candidate_idn = -1  then
           break;
           else
               set e ← candidate_idn
               set cost ← α1.e.T + α2.e. L + α3.e. Q + α4.e. W
               if   V[s_node][i_node] +cost  <  V[s_node][e.next] then
                    set V[s_node][e.next]← cost
           end else
         end for
        set i_node ← mincostNode(V, s_node, directions)
        set directions ← direction_mincostnode(s_node, i_node)
     end while
   end else
return directions
```

2.3. Deadlock freedom

While designing a routing algorithm it is necessary to consider throughput, delay, temperature and energy. Deadlock avoidance is an equally important issue that must be solved. Deadlock refers to a situation in which nodes demand a set of resources and no progress can be made due to cyclic dependency [21]. There are two ways most commonly used to avoid deadlock in NoC.

a. Using virtual channels [22]
b. Some turns in the turn model are not allowed [23]

Using virtual channels leads to higher buffer costs and hardware overhead costs. This work uses the turn model approach to prove that the TADAR algorithm is free from cyclic dependency. For each transmission, depending upon the location of the destination node few turns are prohibited. Consider Figure 2(a) north-west, west-north turns are prohibited in Figure 2(b) south-east and south-west turns are prohibited. Thereby, deadlock freedom is guaranteed due to the elimination of cyclic dependency in the TADAR algorithm.

3. RESULTS AND DISCUSSION

Simulations for the TADAR algorithm have been conducted in a cycle-accurate access Noxim [24] simulator. The simulator is an integration of HotSpot [25] and Noxim [26]. The architectural level thermal model and network model is provided by HotSpot and Noxim, respectively. Overall, access Noxim is capable of generating a network, power, and thermal model for 3D-NoC.

3.1. Simulation setup

In this work an 8×8×4 completely connected 3D-NoC has been taken into consideration to evaluate the performance of the TADAR routing algorithm. Table 1 lists the parameters used in the simulation. TADAR and ATAR are compared under various synthetic traffics. Each simulation runs for 200 Kcycles at a different packet injection rate (PIR). PIR (flits/cycle/node) varies from 0.02 to 0.22 with the interval of 0.02. The distribution of the time interval determines the time at which a packet is injected.

Table 1 Simulation parameters

Parameters	Value
Traffic pattern	Random, shuffle, bit-reversal
Buffer size (flits)	16
Packet size (flits)	2~10
Simulation time (cycles)	200 000
Warm-up time (cycles)	10 000
Network dimension	8×8×4
Packet injection interval	0.02
Packet injection rate (flits/cycle/node)	0.02-0.22

3.2. Performance evaluation

Under different packet injection rates, global average delay is reported in Figures 4(a)-(c) under different synthetic traffic patterns such as bit-reversal, random and shuffle. Graph representing the global average delay for ATAR and TADAR is reported in Figure 4(a). It can be observed that both ATAR and TADAR show similar results between 0.02 to 0.04 injection rate under Bit-reversal traffic however, as injection rate increases, divergence in both algorithms becomes more obvious. It is obvious from analyzing the graph that after 0.04, the global average delay served better in TADAR due to its directional nature as compared to ATAR. TADAR has outperformed ATAR by 26.26% improvement under Bit-reversal traffic.

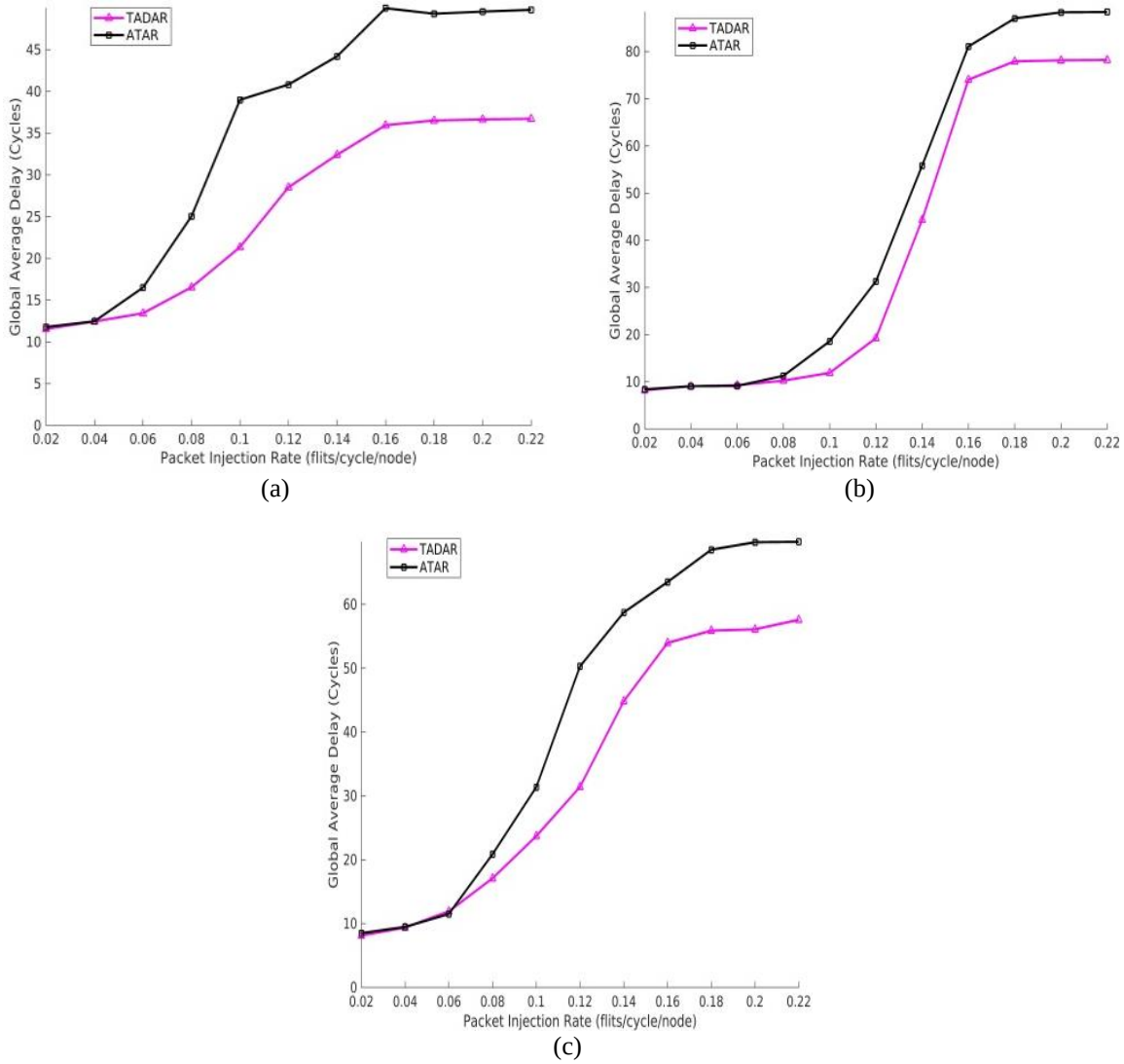

Figure 4. Delay comparison of routing technique under different synthetic traffic patterns (a) bit-reversal traffic, (b) random traffic, and (c) shuffle traffic

Figure 4(b) represents the global average delay of ATAR and TADAR under Random traffic. It can be seen that global average delay is similar between 0.02 to 0.08 injection rates. As the amount of traffic increases both ATAR and TADAR show diverse outcomes. Form 0.08 TADAR has lower global average delay as compared to its counterpart. TADAR survived better due to reaching destinations quickly, reducing the load on the network. Whereas ATAR strives for minimum cost only which forces ATAR to choose unwanted nodes that take away from the destination, hence the causing increase in traffic load in the network. TADAR has achieved a considerable improvement of 11.68% in terms of global average delay under random traffic.

The global average delay under Shuffle traffic between TADAR and ATAR is presented in Figure 4(c). It can be analyzed that injection rates range from 0.02 to 0.06 both routing algorithms show identical results. But on experiencing high injection rates, the difference in performance has begun to emerge. After 0.06, TADAR outperformed ATAR, and overall improvement is recorded at 19.7% under Shuffle traffic. Hence TADAR has performed better than ATAR under all traffic scenarios due to its directional nature resulting in its ability to reach destinations quickly, reducing the overall load on the network. Whereas ATAR strives for minimum cost only to force ATAR to choose unwanted nodes that take away from the destination, hence causing the increase in traffic load in the network.

Considerable reduction in hop counts can be observed in TADAR as compared to ATAR shown in Figure 5. Under Bit-reversal traffic in Figure 5(a), TADAR has considerably 15.5% less hop count as compared to ATAR. Similarly, for Random traffic Figure 5(b), TADAR hop counts are 8% less than that of ATAR and for Shuffle traffic Figure 5(c), 24% improvement reduction in hop count have been observed. Hence TADAR has outperformed ATAR under all traffic scenarios due to its ability to reach its destination quickly reducing the number of overall hop counts. Whereas ATAR strives for minimum cost only to force ATAR to visit unwanted nodes which takes packets away from the destination, hence causing the increase in hop counts under various traffic conditions.

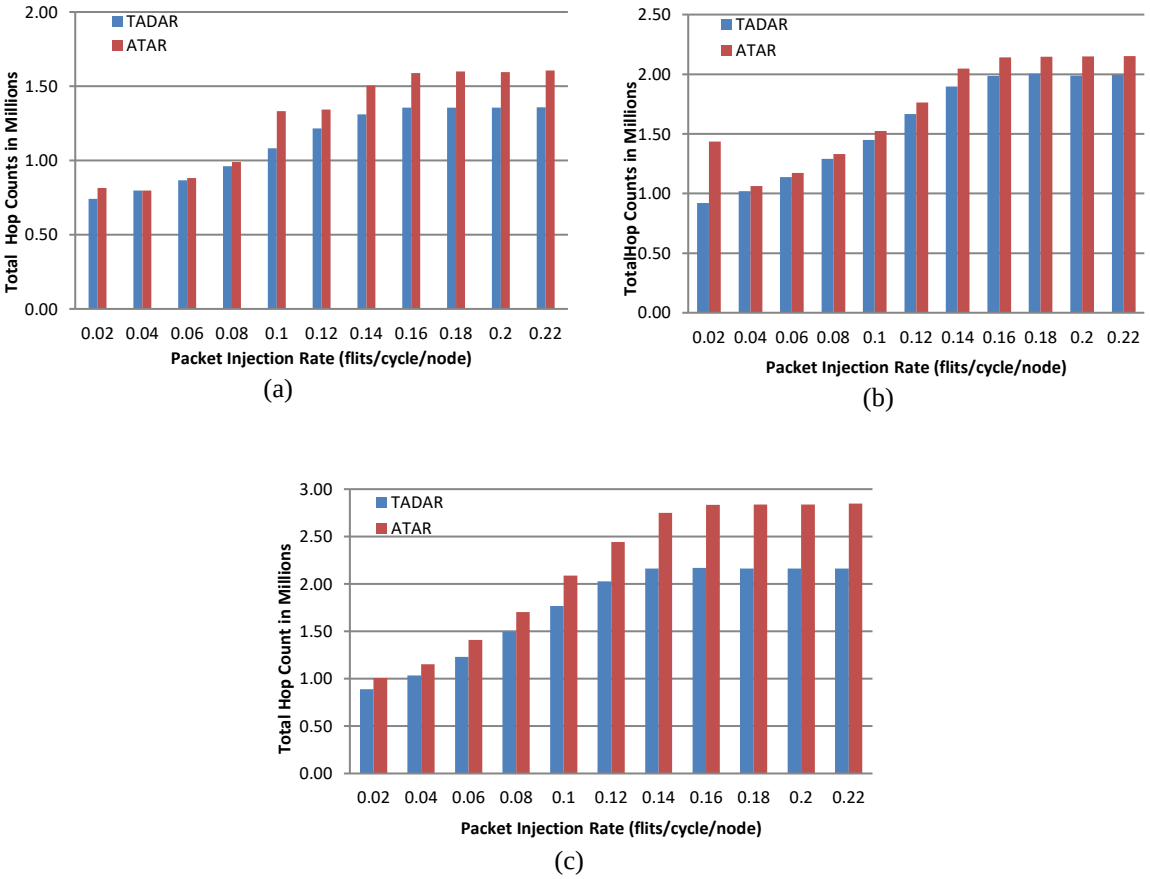

Figure 5. Hop count comparison of routing techniques under different synthetic traffic patterns (a) bit-reversal traffic, (b) random traffic, and (c) shuffle traffic

3.3. Thermal evaluation

Extensive simulations for steady-state temperatures were conducted in this work. Results for thermal profiling at 0.02 PIR for Random traffic are shown in Figure 6(a) and Figure 6(b). During simulations and results comparison, the thermal profile indicates a 5 K decline in peak temperatures of TADAR as compared to ATAR. This indicates that TADAR performed better than ATAR in reducing thermal profiles.

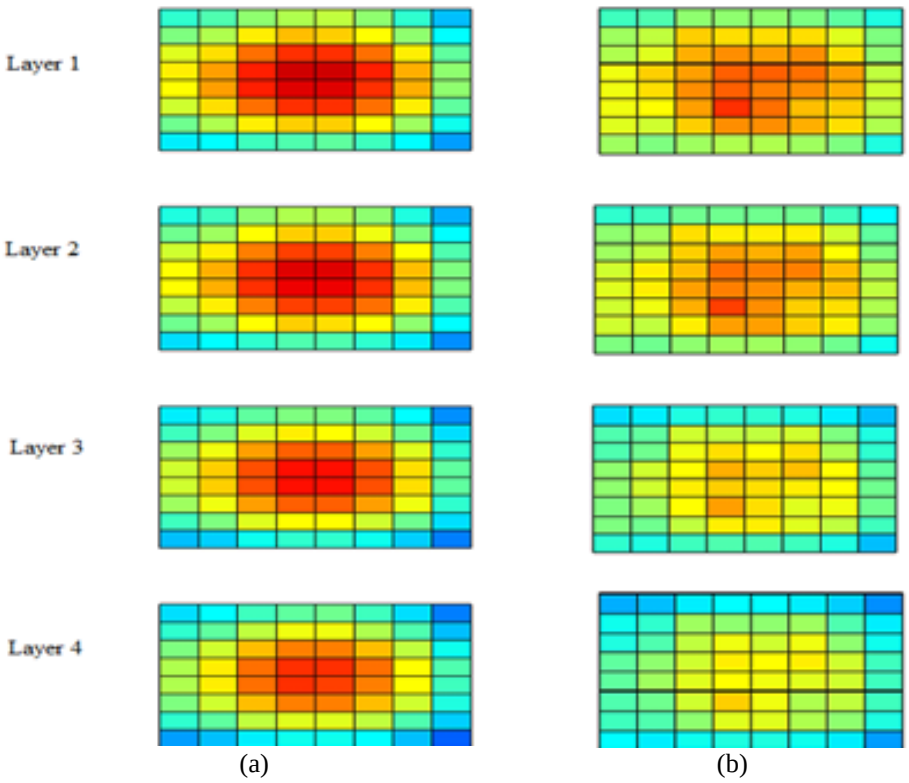

Figure 6. Thermal profile comparison of routing techniques (a) ATAR and (b) TADAR

The simulations have been run several times to ensure that the findings are accurate. Temperature imbalance occurs in ATAR due to longer routes. As a result of the increased traffic, the network becomes congested. Hence, the 3D NoC experiences a higher thermal profile as shown in Figure 6(a). TADAR provides a better thermal profile than other ATAR due to fewer hops, which means packets are reaching their destination quicker and reducing load on network traffic. Hence, fewer packets are consuming buffers and resulting in dissipating less heat. The thermal profile for TADAR is presented in Figure 6(b). At higher PIR, the thermal profile for TADAR is 7 K and 9 K at PIR 0.10 and 0.22, respectively, as compared with PIR 0.02.

4. CONCLUSION

Thermal-aware routing algorithms can reduce thermal hotspots by migrating load from hotter areas of the chip to achieve thermal optimization, resulting in longer paths. In this paper TADAR routing algorithm is presented TADAR is temperature and congestion aware. Most importantly, it strives to find the best possible neighbour to reach closer to the proximity of the destination node. Due to this directional nature, the global average delay has been improved by 11-26%. Similarly, the number of hops in reaching the destination has been reduced approximately 8-24% under various traffic conditions compared to the state-of-the-art routing algorithm. TADAR has shown better results in terms of global average delay, total hop counts and thermal profiles. TADAR has reduced network traffic load by allowing better paths to reach destinations.

ACKNOWLEDGEMENT

The research is supported by Ministry of Higher Education Malaysia (MOHE) and conducted in collaboration with Research Management Center (RMC) at the Universiti Teknologi Malaysia (UTM) under fundamental research grant scheme with grant number: R.J130000.7851.5F029. The authors appreciate greatly for the support.

REFERENCES

[1] K. S. S., A. Jatti, and U. B. V., "Reconfigurable high performance secured noc design using hierarchical agent-based monitoring system," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 8, no. 6, pp. 4164–4174, Dec. 2018, doi: 10.11591/ijece.v8i6.pp4164-4174.

[2] F. W. B. Zulkefli, P. Ehkan, M. N. M. Warip, and N. Y. Phing, "A efficacy of different buffer size on latency of network on chip (NoC)," *Bulletin of Electrical Engineering and Informatics (BEEI)*, vol. 8, no. 2, pp. 438–442, Jun. 2019, doi: 10.11591/eei.v8i2.1422.

[3] A. M. R., A. N. Subrahmanya, and A. D'Souza, "Performance analysis of mesh-based NoC's on routing algorithms," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 8, no. 5, pp. 3368–3373, Oct. 2018, doi: 10.11591/ijece.v8i5.pp3368-3373.

[4] S. Khan, S. Anjum, U. A. Gulzari, T. Umer, and B.-S. Kim, "Bandwidth-constrained multi-objective segmented brute-force algorithm for efficient mapping of embedded applications on NoC architecture," *IEEE Access*, vol. 6, pp. 11242–11254, 2018, doi: 10.1109/ACCESS.2017.2778340.

[5] M. Kaleem and I. F. Isnin, "Thermal-aware dynamic weighted adaptive routing algorithm for 3D network-on-chip," *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 11, pp. 342–348, 2021, doi: 10.14569/IJACSA.2021.0121139.

[6] P. V. Acharya, M. Lokanathan, A. Ouroua, R. Hebner, S. Strank, and V. Bahadur, "Machine learning-based predictions of benefits of high thermal conductivity encapsulation materials for power electronics packaging," *Journal of Electronic Packaging*, vol. 143, no. 4, Dec. 2021, doi: 10.1115/1.4052814.

[7] K. S. S., A. Jatti, and U. B. V., "Design and implementation of secured agent based NoC using shortest path routing algorithm," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 9, no. 2, pp. 950–959, Apr. 2019, doi: 10.11591/ijece.v9i2.pp950-959.

[8] D. Lee, S. Das, and P. P. Pande, "Analyzing power-thermal-performance trade-offs in a high-performance 3D NoC architecture," *Integration*, vol. 65, pp. 282–292, Mar. 2019, doi: 10.1016/j.vlsi.2017.12.002.

[9] D. Lee, S. Das, J. R. Doppa, P. P. Pande, and K. Chakrabarty, "Performance and thermal tradeoffs for energy-efficient monolithic 3D network-on-chip," *ACM Transactions on Design Automation of Electronic Systems*, vol. 23, no. 5, pp. 1–25, Oct. 2018, doi: 10.1145/3223046.

[10] A. Karkar, N. Dahir, T. Mak, and K.-F. Tong, "Thermal and performance efficient on-chip surface-wave communication for many-core systems in dark silicon era," *ACM Journal on Emerging Technologies in Computing Systems*, vol. 18, no. 3, pp. 1–18, Jul. 2022, doi: 10.1145/3501771.

[11] K. Cao, J. Zhou, T. Wei, M. Chen, S. Hu, and K. Li, "A survey of optimization techniques for thermal-aware 3D processors," *Journal of Systems Architecture*, vol. 97, pp. 397–415, Aug. 2019, doi: 10.1016/j.sysarc.2019.01.003.

[12] N. Dahir, A. Karkar, M. Palesi, T. Mak, and A. Yakovlev, "Power density aware application mapping in mesh-based network-on-chip architecture: An evolutionary multi-objective approach," *Integration*, vol. 81, pp. 342–353, 2021, doi: 10.1016/j.vlsi.2021.08.008.

[13] Z. Shirmohammadi, M. Mahmoudi, and M. Rostamnezhad, "Int-TAR: An intelligent thermal-aware routing algorithm for 3D NoC," *Journal of Electrical and Computer Engineering Innovations (JECEI)*, vol. 10, no. 1, pp. 47–56, 2022, doi: 10.22061/JECEI.2021.7750.428.

[14] M. Safari, Z. Shirmohammadi, N. Rohbani, and H. Farbeh, "LETHOR: a thermal-aware proactive routing algorithm for 3D NoCs with less entrance to hot regions," *The Journal of Supercomputing*, vol. 78, no. 6, pp. 1–25, Apr. 2022, doi: 10.1007/s11227-021-04207-3.

[15] N. Rohbani, Z. Shirmohammadi, M. Zare, and S.-G. Miremadi, "LAXY: A location-based aging-resilient Xy-Yx routing algorithm for network on chip," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 36, no. 10, pp. 1725–1738, 2017, doi: 0.1109/TCAD.2017.2648817.

[16] K.-C. J. Chen and Y.-H. Liao, "Adaptive machine learning-based temperature prediction scheme for thermal-aware NoC system," in *2020 IEEE International Symposium on Circuits and Systems (ISCAS)*, Oct. 2020, pp. 1–4, doi: 10.1109/ISCAS45731.2020.9180475.

[17] K.-C. Chen, "Game-based thermal-delay-aware adaptive routing (GTDAR) for temperature-aware 3D network-on-chip systems," *IEEE Transactions on Parallel and Distributed Systems*, vol. 29, no. 9, pp. 2018–2032, Sep. 2018, doi: 10.1109/TPDS.2018.2812164.

[18] S. S. Kumar, A. Zjajo, and R. van Leuken, "Immediate neighborhood temperature adaptive routing for dynamically throttled 3-D networks-on-chip," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 64, no. 7, pp. 782–786, Jul. 2017, doi: 10.1109/TCSII.2015.2503613.

[19] R. Dash, A. Majumdar, V. Pangracious, A. K. Turuk, and J. L. Risco-Martin, "ATAR: An adaptive thermal-aware routing algorithm for 3-D network-on-chip systems," *IEEE Transactions on Components, Packaging and Manufacturing Technology*, vol. 8, no. 12, pp. 2122–2129, Dec. 2018, doi: 10.1109/TCPMT.2018.2842102.

[20] R. Salamat, M. Khayambashi, M. Ebrahimi, and N. Bagherzadeh, "LEAD: An adaptive 3D-noc routing algorithm with queuing-theory based analytical verification," *IEEE Transactions on Computers*, vol. 67, no. 8, pp. 1153–1166, 2018, doi: 10.1109/TC.2018.2801298.

[21] M. Alaei and F. Yazdanpanah, "A high-performance FPGA-based multicrossbar prioritized network-on-chip," *Concurrency and Computation: Practice and Experience*, vol. 33, no. 6, Mar. 2021, doi: 10.1002/cpe.6055.

[22] A. Gangwar, R. Sreedharan, A. Prasad, N. K. Agarwal, and S. H. Gade, "Topology agnostic virtual channel assignment and protocol level deadlock avoidance in a network-on-chipoidance in a Network-on-Chip," in *2021 58th ACM/IEEE Design Automation Conference (DAC)*, Dec. 2021, vol. 2021-Decem, pp. 61–66, doi: 10.1109/DAC18074.2021.9586196.

[23] Jayshree, G. Seetharaman, and D. Pati, "Reliable fault-tolerance routing technique for network-on-chip interconnect," in *Intelligent Sustainable Systems. Lecture Notes in Networks and Systems*, vol. 213, Springer, 2022.
[24] K.-Y. Jheng, C.-H. Chao, H.-Y. Wang, and A.-Y. Wu, "Traffic-thermal mutual-coupling co-simulation platform for three-dimensional network-on-chip," in *Proceedings of 2010 International Symposium on VLSI Design, Automation and Test*, Apr. 2010, pp. 135–138, doi: 10.1109/VDAT.2010.5496709.
[25] W. Huang, S. Ghosh, S. Velusamy, K. Sankaranarayanan, K. Skadron, and M. R. Stan, "HotSpot: A compact thermal modeling methodology for early-stage VLSI design," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 14, no. 5, pp. 501–513, May 2006, doi: 10.1109/TVLSI.2006.876103.
[26] V. Catania, A. Mineo, S. Monteleone, M. Palesi, and D. Patti, "Noxim: An open, extensible and cycle-accurate network on chip simulator," in *2015 IEEE 26th International Conference on Application-specific Systems, Architectures and Processors (ASAP)*, Jul. 2015, vol. 2015-Septe, pp. 162–163, doi: 10.1109/ASAP.2015.7245728.

BIOGRAPHIES OF AUTHORS

Muhammad Kaleem received the MSc degree in computer science from The University of Lahore, Pakistan, in 2017. He is currently working toward the PhD degree in Computer Science at the School of Computing, Faculty of Engineering, Universiti Teknologi Malaysia, Malaysia. His research interests include network-on- chip design, 3D chips, fault-tolerant, and thermal- aware designs. He can be contacted at email: kaleem.muhammad@graduate.utm.my.

Ismail Fauzi Isnin received a M.S. degree in Network Systems Engineering and a Ph.D. degree from the University of Plymouth, U.K., in 2004 and 2011, respectively. Currently, he is a Senior Lecturer in School of Computing, Faculty of Engineering, Universiti Teknologi Malaysia, Malaysia. He is a member of Pervasive Computing Research Group, School of Computing. His research interests are in wired and wireless computers network and communication, mobile ad-hoc network and communication, high performance and parallel computing. He can be contacted at email: ismailfauzi@utm.my.