ROUTING ALGORITHM PERFORMANCE COMPARISON FOR NETWORK-ON-CHIP IN NEUROMORPHIC COMPUTING

[1]JUNRAN PU, [2]VISHNU P. NAMBIAR, [3]WANG LING GOH, [4]ANH TUAN DO

[1,3]School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore
[2,4]Institute of Microelectronics,Agency for Science, Technology and Research (A*STAR), Singapore
E-mail: [1]junran001@e.ntu.edu.sg, [2]ewlgoh@ntu.edu.sg, [3]vishnu paramasivam@ime.a-star.edu.sg, [4]doat@ime.a-star.edu.sg[§]

Abstract - The neuromorphic computing requires a very-large-scale integration system to implement the spiking neural network. However, traditional bus topology is not suitable for network interconnection due to its low scalability. To achieve scalability, network-on-chip is applied in the neuromorphic computing. The routing algorithms used by the network-on-chip affect the networks performance.Unlike the wormhole routing model, which has more than one flit in each packet, the packet of the neuromorphic computing only contains one flit. However, performance analysis of different routing algorithms completed in previous works is based on packets containing multiple flits.In this paper, the performance of the XY, odd-even, and dynamically adaptive and deterministic routing algorithms (DYAD)are analyzed for the neuromorphic computing. The simulation is carried out on the NIRGAM simulator.

Index Terms - Network-on-chip, neuromorphic computing, XY routing algorithm, odd-even routing algorithm, DYAD routing algorithm

I. INTRODUCTION

The human brain has a strong cognitive and learning ability. The neuromorphic computing tries to imitate the human brain to complete the tasks of image and speech recognition. A very-large-scale integration (VLSI) system is used by the neuromorphic computing to achieve the spiking neural network architecture. However,the VLSI system is very complex,and the traditional bus topology causes low communication speed. The network-on-chip (NoC) has been utilized in the VLSI system due to its scalability, flexibility and high communication speed [1]. The routing algorithms used by the router affect the communication performance of the NoC. The oblivious and adaptive algorithms are two major categories of the routing algorithm [2]. The oblivious algorithm does not consider the congestion of the network, whereas the adaptive algorithm does, thereby selecting the less congested route. XY routing algorithm [3], odd-even routing algorithm(OE)[4],and DyAD routing algorithm [5] are three classic routing algorithms that have been widely used in the NoC to route packets. The XY routing algorithm belongs to the oblivious algorithm and it is based on the dimension order. The OE routing algorithm belongs to the adaptive routing algorithm. Comparedwith negative-first, west-first, and north-last [6] routing algorithms, which only allow certain turns to be taken, the OE routing algorithm is more even in nature. The DyAD routing algorithmuses both XY and minimal OE routing algorithms. Previous works [7][8][9][10] have shown the performance comparison of different routing algorithms. However, all of them are based on the wormhole routing model [11], which is not suitable for the neuromorphic computing. In the wormhole routing model, each packet contains more than one flit. For the neuromorphic computing, the information is transmitted through spikes. The spikes are discrete events, so each packet contains only one flit. The TrueNorth[12] developed by IBM uses spikes (packets) that contain the delay time and destination address to transfer information. Each spike (packet) contains only one flit. In this paper, the performance comparison of XY, OE, and DyAD routing algorithms with the packets that contain a single flit is carried out for the neuromorphic computing. Section Ⅱ introduces a basic 2-D mesh NoC architecture. Section Ⅲ gives a brief introduction to three classical routing algorithms. Section Ⅳ shows the experiment results on the performance of three algorithms and compares them. Section Ⅴ draws the conclusion.

II. TWO-DIMENSIONAL MESH NOC ARCHITECTURE

NoC usually consists of the routers and resources. Fig. 1 shows a 3×3 two-dimensional mesh architecture of NoC. The tile represents the resource and R represents the router. Each resource is capable of containing many units, for example,the computation unit, storage unit, network interface unit [1], etc. For neuromorphic computing, the resource is a neuron core, which usually contains a storage unit and a neuron unit. The storage unit stores the synaptic weight values and the neuron unit performs the computation [12]. Each neuron core generates spikes (packets) that are transmitted to other neuron cores through the router. Each spike contains the information of the target neuron core address.These spikes are typically sparse, but spiking activity may suddenly surge at certain periods of the computation stage, depending on the neuromorphic architecture and the intended target application. The routers of the NoC are used to connect the resources and transmit the data from one resource to other resources. Two one-directional buses

are utilized to connect two routers. Each router mainly consists of five ports, five input buffers, an arbiter, and a crossbar switch. Five ports represent five directions: north, east, south, west, and local. The network interface of the resource is connected to the router. The input buffer stores the data arriving at the input port. The arbiter selects the output port for the input data according to the routing algorithm.

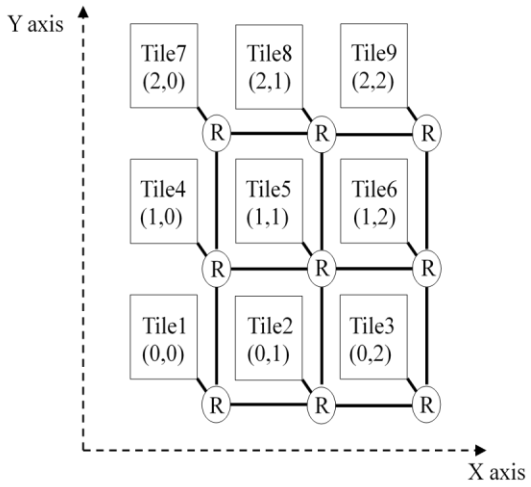


Fig.1. 3×3 two-dimensional mesh NoC architecture.

III. CLASSIC ROUTING ALGORITHM

A. XY Routing Algorithm

The XY routing algorithm is the most widely used oblivious algorithm and this routing algorithm can prevent deadlock and livelock [13]. During the transmission process, the packets occupy the resources of the router, and the resources are released only after the packets get the resources of the next router. When these packets occupy each other's resources, they are not able to obtain the new resources of the next router and release the resources they have occupied. Once a waiting cycle is formed, the deadlock occurs. Livelocksurfaces when the packets are unable to reach the destination, and this happens in the adaptive and nonminimal routing [2]. The XY routing algorithm routespackets on the basis ofdimension order. The x-direction has the highest priority, and the packets will only move in the y-direction when the packets reach the same x coordinate as the destination address. For example, tile 1 in Fig. 1 generates a packet and the destination of this packet is tile 9. The packet is first transmitted in the x-direction, from tile1 to tile2 and then to tile 3. The x coordinate of tile 3 is the same as tile 9.Thereafter, the packet is transmitted in the y-direction, from tile3 to tile6 and finally,arrives at tile 9.The XY routing algorithm is very easy to implement in the router design. However, the routing path of this algorithm is fixed once the destination is determined. Besides, XY routing algorithm will cause uneven distribution of traffic. Hence, it canlead to congestion in the middle of the network [2]. Having said that, the XY routing algorithm can be readily modified. Many new algorithms, such as the pseudo adaptive XY routing [13], dynamic XY (DyXY) routing [14], location-based aging-resilient Xy-Yx routing algorithm (LAXY) [15] are all created based on the XY routing algorithm.

B. Odd-Even Routing Algorithm

The OE algorithm is an adaptive routing algorithm where no virtual channel is needed in this algorithm. Two rules are used in the OE routing algorithm to restrict certain turns to be taken based on the location. The OE routing algorithm is deadlock-free. Besides, the probability of hot spot traffic occurring in this algorithm is reducedtremendously [4].

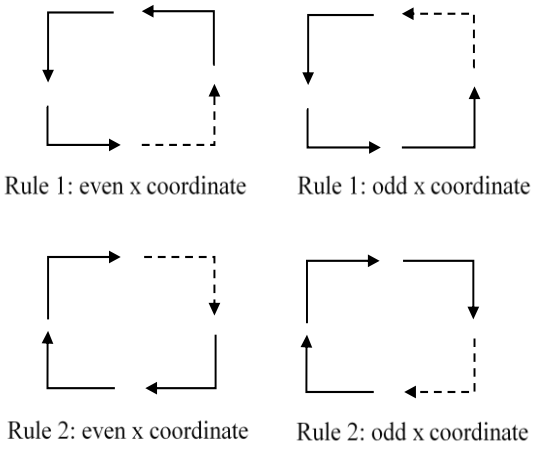


Fig.2. Allowed turns for rule 1 and rule 2.

The position of a router is represented by (x,y). In rule 1, packets are not allowed to change the direction from east to north if the x coordinate is an even number.If the x coordinate is an odd number, the packets are not allowed to change the direction from north to west. As for rule 2, packets are not allowed to change the direction from east to south if the x coordinate is an even number. If the x coordinate is an odd number,the packets are not allowed to change the direction from south to west [4]. Fig. 2 shows the turns that are allowed for rule 1 and rule 2. The solid line represents the allowed turns and the dotted line represents the prohibited turns. The north-last, negative-first, and west-first routing algorithms proposed in [6] are partially adaptive. Each of these algorithms prohibits two turns to prevent the deadlock. However, these three partially adaptive routing algorithms cause unfairness and network congestion. Comparing betweenthe three algorithms, the OE routing algorithm is more even and is capable of achieving higher routing adaptiveness [4].

C. DyAD Routing Algorithm

The DyAD algorithm combines the XY and OE routing algorithms. The XY routing algorithm is applied when the network is not busy, and the minimal OE algorithm is engaged if the network is congested. The minimal OE routing algorithm is used to prevent the livelock. The network congestion

situationdepends on the congestion threshold value. If thecorresponding input FIFOoccupation ratio of the neighboring router is higher than the congestion threshold value, this means the neighboring router is in the congestion situation, and the minimal OE algorithm should be utilized to avoid the congestion. Otherwise, the neighboring router is not congested, only XY routing algorithm is required. No virtual channel is used in this algorithm [5].

The advantage of the XY routing algorithm is that it can achieve lower routing latency than the OE algorithm,but the XY algorithm canlead to the network congestion. The OE routing algorithm can reduce network congestion and achieve larger throughput than the XY routing algorithm, but it results in higher communication latency [5]. Thus, the DyAD routing algorithm combines the advantages of these two algorithms to achieve greaterrouter performance.The DyAD algorithm can achieve the same latency as the XY algorithm when the network is not in the congestion state. Besides, the DyAD routing algorithm can also achieve high throughput.

Table IParameters of the NIRGAM Simulator

Parameter Name	Value
Topology	Mesh
Network Size	3×3
Routing Algorithms	XY, OE, DyAD
Number of Buffers	16
Packet Size	4 bytes
Clock Frequency	1GHz
Simulation Clock Cycles	20000
Warm up Cycles	500
Packets Generation Cycles	1000
Flit Interval	CBR: 2 clock cycles
	BUR: 1 clock cycle
Load Percentage	10%~90%
Average Burst Length	10
Average Off Time	4

IV. SIMULATION RESULTS AND PERFORMANCE EVALUATION

The NIRGAM simulator [16] is used in this experiment to analyze the performance of XY, OE, and DyAD routing algorithms. NIRGAM is a SystemC based simulator for performance analysis of the NoC. In this experiment, each packet only consists of one flit. Latency represents the total time it takes to deliver the packet from source to destination. Throughput is defined asthetotal number of flits each router can deliver per cycle.Total network power is defined as the total network power consumption.The percentage of maximum bandwidth usage depends on the value of the load. Constant bit rate (CBR) traffic generator and bursty (BUR) traffic generator are used to generate the packets. The average burst length of BUR traffic generator represents the number of packets generated within a given time. The average off time represents the time interval between the bursts. The experimental setups of the NIRGAM simulator are shown in Table I.

Power consumption, latency, and throughput comparison of XY, OE, and DyAD algorithms are carried out in this experiment. Fig. 3 shows the simulation results of the total network power (in milliwatts) of XY, OE, and DyAD algorithms. The load percentage varies from 10% to 90% in 10% interval. In general, when the load percentage is lower than 50%, the total network power of these three algorithms increases with the increase of the load percentage. When the load percentage is greater than 50%, the total network power of these threealgorithms changes a little as the load percentage increases. The OE algorithm consumes higher total network power as compared to other two routing algorithms. The total network power difference between the XY and DyAD routing algorithm is small. Compared with the CBR traffic generator, the BUR traffic generator consumes less total network power.

Fig. 4 shows the overall average latency per channel of XY, OE, and DyAD algorithms. The OE algorithm has the highest average latency per channel among the three routing algorithms, whilethe DyAD algorithmachieves the lowest average latency. For these three algorithms, as the loadpercentage increases, the average latency per channel does not change significantly. Compared to the CBR traffic generator,the BURtraffic generator has higher latency per channel.

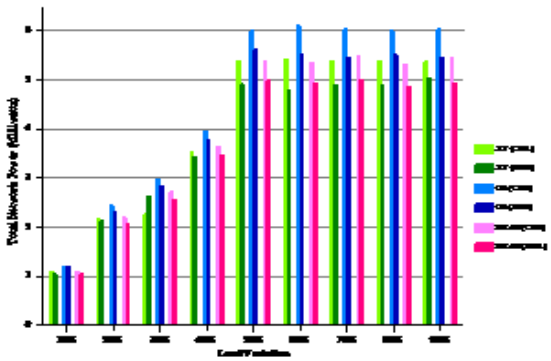


Fig.3. Total network power.

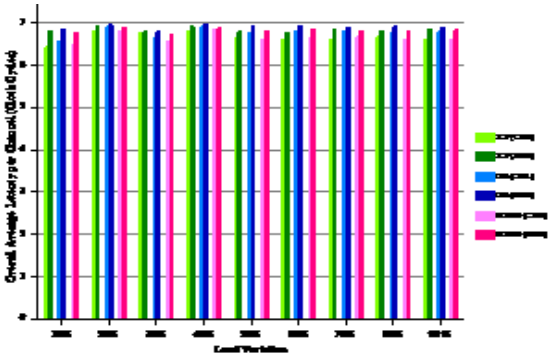


Fig.4. Overall average latency per channel.

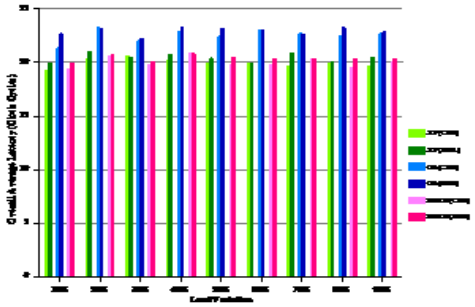


Fig.5. Overall average latency.

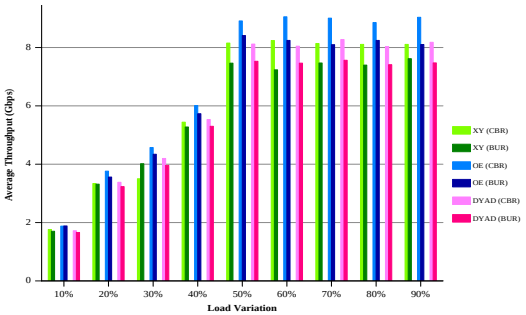


Fig.6. Average throughputper channel.

Fig. 5 shows the overall average latency of XY, OE, and DYAD routing algorithms. The OE algorithm has the highest average latency among the three algorithms. Compared to the BUR traffic generator, the CBR traffic generator has lower overall average latency. Fig. 6 shows the average throughput per channel of XY, OE, and DyAD routing algorithms. In general, when the load percentage is lower than 50%, the average throughput of these three routing algorithms increases with the increase of the load percentage. When the load percentage is higher than 50%, the average throughput of these three routing algorithm does not change significantly as the load percentage increases. The OE routing algorithm has the highest average throughput among the three algorithms. The average throughput of DyAD algorithm is slightly higher than the XY algorithm. The CBR traffic generator has higher average throughput as compared to the BUR traffic generator.

CONCLUSION

In this paper, performance evaluation of XY, OE, and DYAD routing algorithms is carried out on the NIRGAM simulator. In this experiment, each packet only contains one flit, which is the same as the neuromorphic computing. Thus, these experimental results provide a reference for the neuromorphic computing when selecting routing algorithms. The network size is 3×3 in this experiment. From the simulation results, it can be seen that when the load percentage is less than 50%, as the load percentage increases, the network average throughput and total network power increase. When the load percentage is greater than 50%, the network average throughput is saturated, and the total network power does not change significantly. The OE routing algorithm has the highest average throughput and total network power among the three algorithms. Compared with the CBR traffic generator, the BUR traffic generator consumes less total network power and has lower average throughput. For the network latency, the OE routing algorithm has the highest overall average latency among the three algorithms. The CBR traffic generator has lower overall average latency when compared to the BUR traffic generator.

REFERENCES

[1] S. Kumar et al., "A network on chip architecture and design methodology," in VLSI, 2002. Proceedings. IEEE Computer Society Annual Symposium on, 2002, pp. 117-124: IEEE.

[2] V. Rantala, T. Lehtonen, and J. Plosila, Network on chip routing algorithms. Citeseer, 2006.

[3] A. Intel, "Touchstone DELTA system description," Supercomputer Systems Division, Intel Corporation, Beaverton, OR, vol. 97006, 1991.

[4] G.-M. Chiu, "The odd-even turn model for adaptive routing," IEEE Transactions on parallel and distributed systems, vol. 11, no. 7, pp. 729-738, 2000.

[5] J. Hu and R. Marculescu, "DyAD: smart routing for networks-on-chip," in Proceedings of the 41st annual Design Automation Conference, 2004, pp. 260-263: ACM.

[6] C. J. Glass and L. M. Ni, "The turn model for adaptive routing," ACM SIGARCH Computer Architecture News, vol. 20, no. 2, pp. 278-287, 1992.

[7] P. Parandkar, J. Dalal, and S. Katiyal, "Performance Comparison of XY, OE and DY Ad Routing Algorithm by Load Variation Analysis of 2-Dimensional Mesh Topology Based Network-on-Chip," BIJIT Journal, vol. 4, no. 1, pp. 391-396, 2012.

[8] J. K. Singh, A. K. Swain, T. N. K. Reddy, and K. K. Mahapatra, "Performance evalulation of different routing algorithms in Network on Chip," in Microelectronics and Electronics (PrimeAsia), 2013 IEEE Asia Pacific Conference on Postgraduate Research in, 2013, pp. 180-185: IEEE.

[9] N. Soni and K. Deshmukh, "Comparison between three different types of routing algorithms of Network on Chip," in Advances in Optical Science and Engineering: Springer, 2015, pp. 447-459.

[10] N. Y. Phing, M. M. Warip, P. Ehkan, F. Zulkefli, and R. B. Ahmad, "Performance Analysis of the Impact of Design Parameters to Network-on-Chip (NoC) Architecture," in International Conference of Reliable Information and Communication Technology, 2017, pp. 237-246: Springer.

[11] L. M. Ni and P. K. McKinley, "A survey of wormhole routing techniques in direct networks," Computer, no. 2, pp. 62-76, 1993.

[12] P. A. Merolla et al., "A million spiking-neuron integrated circuit with a scalable communication network and interface," Science, vol. 345, no. 6197, pp. 668-673, 2014.

[13] M. Dehyadgari, M. Nickray, A. Afzali-Kusha, and Z. Navabi, "Evaluation of pseudo adaptive XY routing using an object oriented model for NoC," in Microelectronics, 2005. ICM 2005. The 17th International Conference on, 2005, p. 5 pp.: IEEE.

[14] M. Li, Q.-A. Zeng, and W.-B. Jone, "DyXY: a proximity congestion-aware deadlock-free dynamic routing method for network on chip," in Proceedings of the 43rd annual Design Automation Conference, 2006, pp. 849-852: ACM.

[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.

[16] L. Jain, B. Al-Hashimi, M. Gaur, V. Laxmi, and A. Narayanan, "NIRGAM: a simulator for NoC interconnect routing and application modeling," in Design, Automation and Test in Europe Conference, 2007, pp. 16-20.