An Efficient Directional Routing Algorithm For Network On Chip

Venkateswara Rao Musala, Venkata Rama Krishna Tottempudi

Abstract: Bus structures are commonly used in System on Chip (SoC) which needs a lot of wiring that causes an increase in Resistance and Capacitance (RC) of the framework in SoC. To avoid this an interconnection network called Network on Chip (NoC) is introduced for better communication in terms of latency and throughput among the processing cores in the vicinity of the selected network. It plays a major role to dress the issues in SoC. An on-chip routing resource is used to send the data packet based on routing decisions done in the router, which improves performance of interconnection fabric in terms of latency and throughput over resolute wiring and buses. Present routing algorithms in NoC experience a problem of channel load imbalance, which causes congestion in the routed path and effects the latency and throughput of the routed packet. This work proposes an adaptive routing resource fabric (Directional Routing Algorithm (DRA)) to avoid the congestive paths by identifying the unloaded path with the help of timeout piggybacking and load shedding, the DRA bypasses the congested path on the channel, based on direction specific traffic patterns. The proposed algorithm does better than Normal XY routing by 18% and 31% in terms of Avg.latency and throughput

Index Terms: Avg. Latency, Directional Routing Algorithm (DRA), Network-on-Chip (NoC), Resistance and Capacitance (RC), System on chip (SoC), Throughput .

1. INTRODUCTION

Various processing cores and memory cores are integrated on a single chip to perform the different functionalities of application specific SoC. Rantala, et al. [1], presented two significance confinements in SoC: the first significance, Intellectual Property (IP) based communication drops the perfection at the architecture and increase the wiring delay. Second issue is it continuously consolidates many processor for different applications on the same chip, because of these confinements, no actual structure for combined SoC application, the size of the components scales down [2], in NoC it is required to minimize the application complexity and bandwidth, channel latency makes an interpretation of the processor to attain high performance in future multi-core processor architectures. There are numerous ways are available ways are available to communicate among the Source Processing Core (SPC) and Destination Processing Core (DPC) in interconnection network. An algorithm is used to route the data from SPC-DPC, present path routing resources are concentrated on oblivious routing resources like Dimension-Order-Routing (DOR), route the packets regardless of the load among the routed paths, however these processes [4] have less multifaceted nature and exhibits poor communication metrics. An adaptive algorithm needed to route the data packet through the less congested path by considering channel loads at each Source to Destination (SX-DX) pair. In this aspect Carrying out a routing algorithm with adaptiveness is important, routing resource [1] shows a major role in the selection of resource based on the interconnections in the network, in terms of latency, throughput, energy and area. In this proposed work we considered latency and throughput as the design metricsThe Fig.1 demonstrate the Data packet (DP) communication model for the communication between the routers, before sending the data packet to the destination router the data packet should undergo the router one, although the base router infer with congestion the source router must route the packet through the same as a result of it follows the XY routing. The limitations for congestion in the routed path are router contention, link congestion and router contention

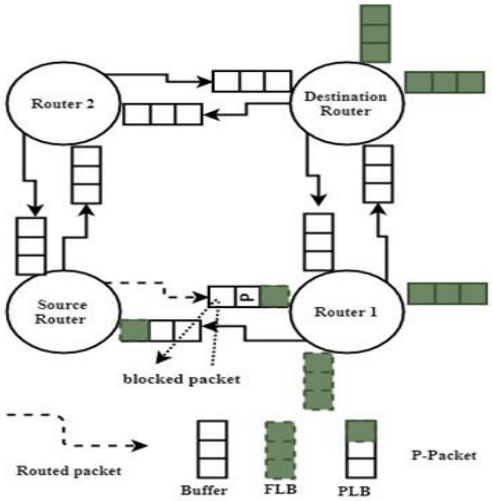

***Fig.1**. Router communication model*

Congestion of routed Packet starts with the blockage of data traffic these kind of blocks are happen because of deadlock, live-lock and starvation [1] so the data is must route through the data links without any of the above said routing limitations for this in this paper we considered the channel congestion as the contention metric, to identify the channel congestion rate we used a timely piggybacking to get the acknowledgement of packet by the neighbour node.at same time load shedding techniques is used to balance the load in the routed channel of the packet Congestion [5-8] causes when source node and the adjacent node want to use the same routing path to send the packet to the other nodes in the architecture and the bandwidth lies in the specified output port format, defined by limitations of the network fabric. The input traffic is sent irregularly from the source node among output ports. It is easy to imagine a situation during which network limits with full traffic intensity are prodigious [9]. Quadratic implementation is used to design the network topologies. In this processing nodes are arranged in dimensional view. In the literature majority of the researchers are used network topologies such as mesh, torus etc. to transfer the data, which enlighten that mesh outperforms best than other topologies The routing

resource is crucial to transfer the data packet in the network, X-Y routing may be used to overcome data blockage between SX-DX sets and channel selection changes the limitations (latency, network traffic and etc.) of the network taken into consideration [5].

2 PREVIOUS WORK

W.J.Dally et.al [10], Introduce an algorithm Globally Oblivious Adaptive Locally (GOAL) for torus networks to minimise the fully randomizes routing and to increase the throughput of the communication channel, the GOAL is compared with the other routing algorithms, this algorithm will work only for k-ary n-cube (Torus) networks. Schwiebert and Bell [11], described work done on adaptive routing algorithms using worm-hole(WH) routing the advantages of WH is optimization of channel latency by sending the head flit immediately based on the availability of the adjacent processing element, drawback in WH is susceptible to contention problems. Huang and Hwang [12], given many insights over the routing algorithms used and future scope in NoC. Nilsson, et al. [13], introduced a Proximity Congestion Awareness (PCA), technique to calculate the congestion metric by using FIFO, this work is limited to random traffic only. Oommen and Jose [14], introduced a NoC Router with adaptive routing which provides more than one possible paths to incoming data packet based on the routing decisions done by the routing algorithm, adaptivity in the routers will take care of turn models and odd-even models. Hu and Marculescu [15], introduced a routing techniques to switch among the deterministic and adaptive routing based on the congestion levels in the router, the mode of switching to either deterministic or adaptive need to be done for effective switching strategies under different traffic patterns. Pande, et al. [8], introduced a regular and significant assessment procedure to compare the performance and characteristic parameters of existing NoC topologies. Inclusion of parameters such as testability, dependability, and reliability is an added security aspect for future NoC architectures. Gratz, et al. [16] Introduced a Regional congestion Algorithm (RCA) to improve load balance in interconnection networks but it is limited to mesh topology under minimal routing techniques further RCA can be extended to non-minimal adaptive routing by accounting the global contention. Wang, et al. [17], The congestion-conscious router was implemented by taking the amount of service routes in the network as an Index of traffic load balancing since Dmesh is capable of delivering better-integrated services and of tolerating failures. Kim, et al. [18], Used two pipeline strategies to reduce inter-connection network latency, but to enhance the use of the network router area will increase as per M/G/1/M queue latency model is defined, [Kim, et al. [18]][Kim, et al. [18]] Kiasari, et al. [19], introduced G/G/1 model to analyse the latency of the channel. Chang, et al. [3], used congestion metric to calculate the contention in the network. Aswathy, et al. [19], introduced a packet regulating mechanism to avoid the congestion the technique used by the authors is increases the serialization packet delay in NoC. In this case the oblivious routing is done without channel traffic, while adaptive routing takes into consideration both communication channel, the contention of the router and prevents the congestion of most of the last practical latency in the NoC architectures[1].Several networks in NoC are developed to cre ate specific topology with the addition of traffic patterns.

Chang, et al. [3] Used a FIFO method to improve efficiency and resolve the allotment of buffer space based on actual data packet dimensions in NoC-based systems to perform communication among S and D nodes, however, strategies cannot manage the wormhole-switched networks. The model described in Singh, et al. [20], applicable for single buffer networks and does not consider queue-up delays and network disputes which Introduce the link capacity allocation in the NoCs using a serialization analysis latency model. the work presented [15] describes the analytical routing The prior works assume Poisson distribution as the dataset f or injection of data into the flow of information packets, such models in many applications do not have the precise replacing of congested traffic pattern implementations.Jafari, et al. [21] A latency flow study is defined for pseudo-uniform networks,Rohbani, et al. [22], Explained the integrated traffic control and network channel prevention metrics but this strategy was not perfect for random transportation of such a system with real-time demands.

3 METHODOLOGY OF THE PROPOSED WORK

In NoC, the resource of routing is used to minimize the network latency. In this proposed work, an improved routing algorithm is used to monitor network load, throughput and serialization latency in the communication channels when non-uniform traffic is included [11]. Surprisingly many built-in routers use a pattern such as traffic transformation and arbitrary traffic [20], make a pitiful task of load balance. On a miserable, pre-determined route the traffic between every couple of nodes. As laid down by [5, 18, 23], the non-uniform model of traffic can lead to significant load imbalances and less throughput within the network. The method implemented in this work is used to get acknowle dgment from the beneficiary node based on the time out frame work designed in the router, if there is no acknowledgment in specific time the proposed heuristic is provides an alternative pathway from SX-DX in successive clock cycle to chop the packet waiting time in routes with the help of technique used in [24] and minimizes the route allocation latency, the load shedding method is prioritize the data to be sent as packets according to the recipient node. This is the first network model proposed to calculate the Avg.latency of the network. this model often creates efficient performance analysis under random traffic patterns with wormhole switching [2] using synthetic data . In addition to offering network trade-offs, our suggested model improve network latency and application mapping.

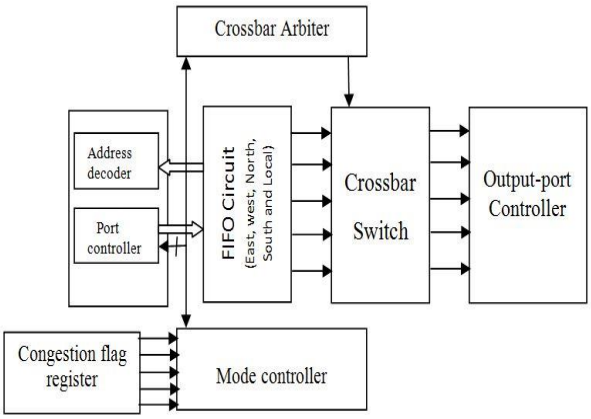

***Fig.2**. Block diagram of proposed router*

In interconnection networks i. Oblivious Routing (OR) and ii.

Adaptive Routing (AR) is popular, OR specifies the path from SX-DX based on possible and optional paths in OR the information packet is transmitted according to the topology chosen for this type of network. of singularity causes congestion in the routed path the literature [25] address this problem by minimizing the traffic , whereas AR [1] selects the path adaptively in this proposed work we adopted the AR with XY routing technique [26, 27] to select the pre-eminent path by avoiding the livelocks. The routing decisions in the proposed algorithm are simple and less time complex with better reliability at the same time it is having more advantageous to NoCs [28].

4 FUNCTIONAL MODEL OF DRA

Well-Designed routing algorithms optimally select the length of the communication channel, decrease the size of the hop, decrease the general latency of serialization, balance load as well as optimize performance. In reality, we have to raise the average duration of all information packs in order to enhance the general load balance in case of overlooked routing algorithms. The opposite is true, too. This compromise occurs for ORs because they do not influence the present pattern of traffic in the routing scheme. The ability to function in the context of network deficiencies is also an important component of a routing scheme. If the network flops to establish the communication among SX-DX, then whole system may fail, although the algorithm can be programmed or adapted to failure, then only a slight loss of efficiency can keep the system functional. This is clearly critical for highly reliable systems. At last, routing synergizes with the network load balancing and careful design is often needed to prevent deadlock and/or livelocksFig.3 describes the mesh topology with 36 processing cores or Elements (PE), Topology is organized, any PE in the network can be used as a source or Destination the communication in the topology is done using the outing scheme specified in Fig.1. The aim RA is to provide better latency, throughput, power, and reliability.

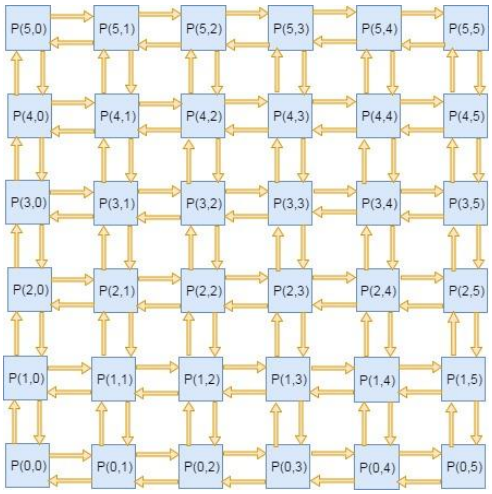

***Fig.3**. General Mesh topology with 36 PEs*

The DRA is used to calculate the congestion metric and cost of communication with the help of the figure shown in Fig.4. A load shedding method is used to prevent congestion of all feasible routes from SX to DX [16] and timeout piggybacking [21] decrease transaction of packet time throughout the router

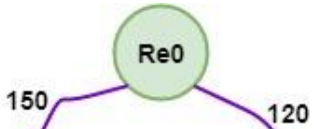

Fig.4. Resource Core Graph (RCG)

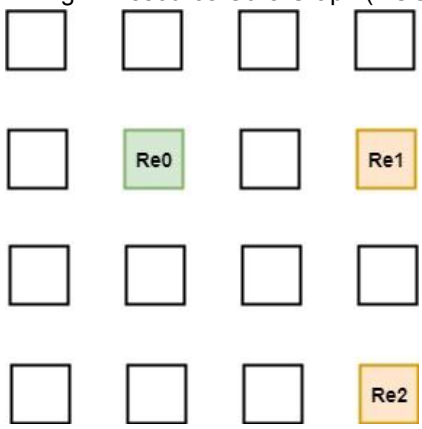

***Fig.5**. Placing of resources in 4X4 NoC*

Re0=(x1, y1) = (2, 1), Re1=(x2, y2) = (2, 3)

Re2=(x2, y2) = (0, 3)

$$Communication\ Cost = Communication\ bandwidth * Communication\ distance$$

$$Communication\ \ distance\ between\ two\ nodes = (|x2 - x1| + |\,y2 - y1|)$$

$$Communication\ cost\ from\ Re0\ \rightarrow Re1 = 150 * (|2 - 2| + |3 - 1|)$$

= 300

$$Communication\ cost\ from\ Re0 \rightarrow Re2 = 120 * (|0 - 2| + |3 - 1|)$$

= 480

$$\text{Total } Communication\ cost = 780$$

In the proposed algorithm the data packets are sent via the minimized load channel. Queue length is used to approximate the load of the serving channel, a number of packets transferred from SPC to DPC completed within the time (T) slots based on the communication cost calculation. All these are applied at SPC to minimize the router and channel latency. Subsequent analysis demonstrates the impact of load balance on routing control and the detection of defective nodes by DRA algorithms considerably in Fig. 6 The SPC is PC (0 1) and the DPC is PC (3 3), the selection of the routing path for the proposed algorithm will choose based on the cost for communication from SPC to DPC.

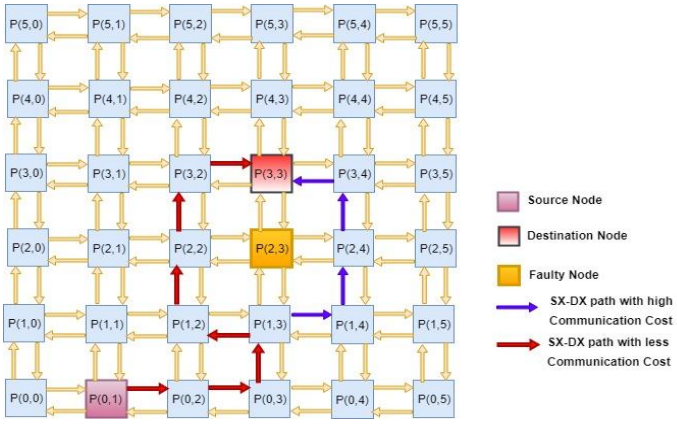

***Fig.6**. 6X6 mesh topology with proposed model*

The possible pathways (Pw) from SPC to DPC is as follows

$$Pw1 = PC(0,1) \rightarrow PC(0,2) \rightarrow PC(0,3) \rightarrow PC(1,3) \rightarrow P(2,3) \rightarrow PC(3,3)$$

$$\mathrm{Pw2} = \mathrm{PC}(0,1) \rightarrow \mathrm{PC}(0,2) \rightarrow \mathrm{PC}(0,3) \rightarrow \mathrm{PC}(1,3) \rightarrow \mathrm{PC}(1,2) \rightarrow \mathrm{PC}(2,2) \rightarrow \mathrm{PC}(3,2) \rightarrow PC(3,3)$$

$$Pw3 = \mathrm{PC}(0,1) \rightarrow \mathrm{PC}(0,2 \rightarrow \mathrm{PC}(0,3) \rightarrow \mathrm{PC}(1,3) \rightarrow \mathrm{PC}(1,4) \rightarrow \mathrm{PC}(2,4) \rightarrow \mathrm{PC}(3,4) \rightarrow \mathrm{PC}(3,3)$$

DRA uses piggybacking as a timing index to identify the faulty nodes in the communication path from SN to DN. From the above three possible paths are exist from SPC to DPC, in the possible pathways the communication through Pw1 is not permissible because the node PC (2, 3) is a faulty node, it is identified by using piggy backing in this technique to form the communication between SPC and adjacent node a request is sent from the SPC in X direction is there any acknowledgement from the adjacent receiving node then the communication is established from the adjacent node which form the communication path based on the routing algorithm instructions, When no accusations are received, then the SPC sends the request in –X direction to the neighboring node and the Y direction from the –X direction destination. SPC calculates communication cost (CC) by the space among SPC, DPC. Which is calculated from two different routes from the SPC to DPC Pathway-2(Pw2) and Pathway-3(Pw3),

From the figure 4 &5 the communication cost is calculated as follows

The CC of Pathway2 is CC-Pw2=29 and pathway3 is CC-Pw3 =34. Due to the communication expenses of the routes, SPC chooses the Pw2 for PC (0, 1) to PC (3, 3) communication. Calculation of communication cost is done as shown in Fig. 4 & 5.

5 RESULTS & DISCUSSIONS

i Simulation Tool

In this work a precise network-on-chip-based simulator oriented on C++ [29] is used to assess the performance of the topology with various buffer depths and traffic patterns, FIFO arbitration and wormhole switches are used to reduce the serialization latency during arbitration. Single cycle based flit transfer is used during the transmission of data packet to minimize the congestion among the channel and router, the data packet may have flits ranges from 0-9 flits including the head and tail flits. For simulation 10,000 cycles are used and 1000 cycles are used for warm-up.

ii Traffic patterns

In this proposed work various traffic scenarios, such as uniform and random are used to test the DRA algorithm and the results are compared with [22, 28] the findings show a better latency and throughput of the proposed algorithm.

iii Evaluation Metrics

Throughput and Avg. Latency are the metrics [21, 23] to evaluate the performance of the proposed work. Latency is a quality metric to assess the travel time of the data packet via network including data packet injection into the channel, channel wait time and the receipt of tail flit at DPC. Whereas throughput is another metric to assess the quantity of information is received by the DPC in (bit / sec).

iv Step wise execution of DRA

Step1: Identify the Source node and Destination node
Step2: Identify the blockage node in the routed path
Step3: Compare the coordinates of Source and destination nodes
Step4: If source node is greater than destination node decrement coordinate of the source
Step5: repeatedly compare the updated source node with the destination node
Step6: repeat step4 for another coordinate
Step7: increment the count based on the step4 and step6 until packet reaches to the destination

The results indicates that DRA routing achieves ensemble performance associated with the conventional XY routing algorithm, the simulation results are checked with various data packet injection rates from 0.2 packets/cycle to1 packets/cycle on 16x16 mesh topology with 64-bit flit size for various patterns such as random and homogeneous with the aid of XY and proposed DRA algorithm. The data packet transferring from SPC through the link formed by the DRA as head, body, and tail flits. The CC between PE's is optimised using availability of adjacent nodes. If the neighboring node is ready to serve SN, then both the acknowledgment and data is sent at the same time which reduces CC between PE's to an optimal point, but which adds the area overhead. VALID and WAIT control signals are used to know the status of the data packet. The DRA routing algorithm's Avg. latency is contrasted with the traditional XY routing algorithm with distinct buffer dimensions of 2, 4 in random patterns and consistent patterns of traffic. The results of the DRA algorithm are shown belowAvg.latency of the mesh topology under random traffic patterns

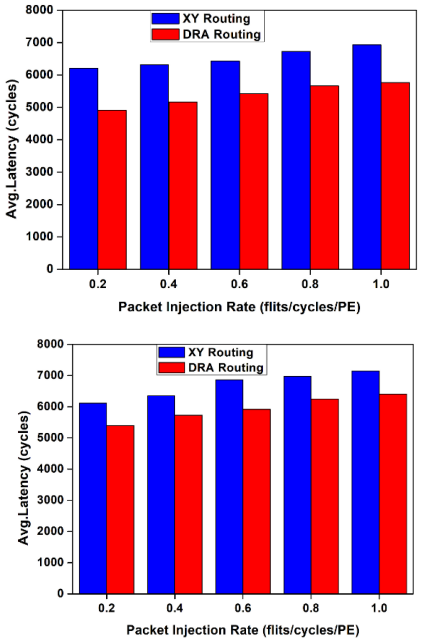

***Fig.7**. Latency vs. Packet Injection Rate with buffer size=2, 4*

Avg.latency of the mesh topology under uniform traffic patterns

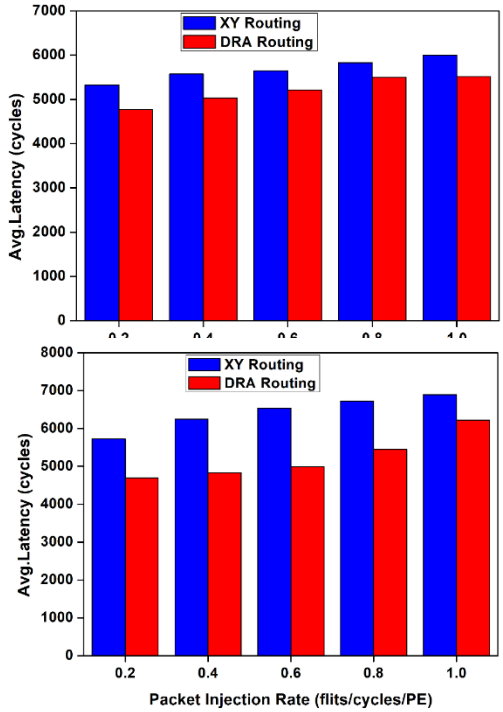

***Fig.8**. Latency vs. Packet Injection Rate with buffer size=2, 4*

Throughput of the mesh topology under random traffic patterns

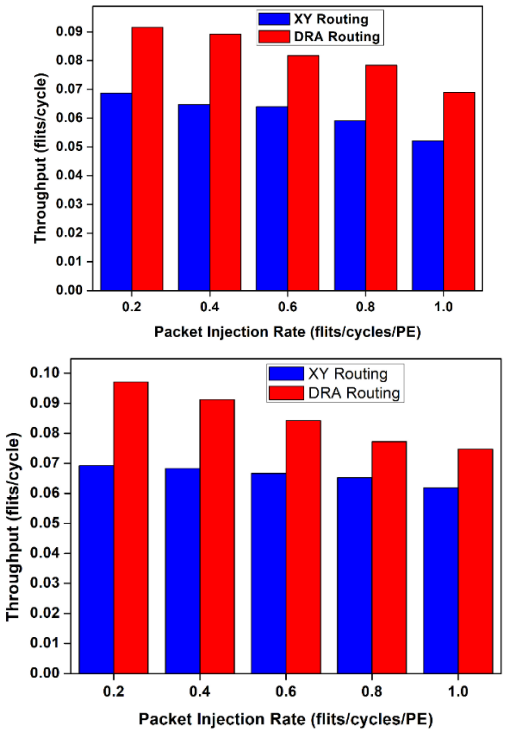

***Fig.9**. Throughput vs. Packet Injection Rate with buffer size=2, 4*

Throughput of the mesh topology under uniform traffic patterns

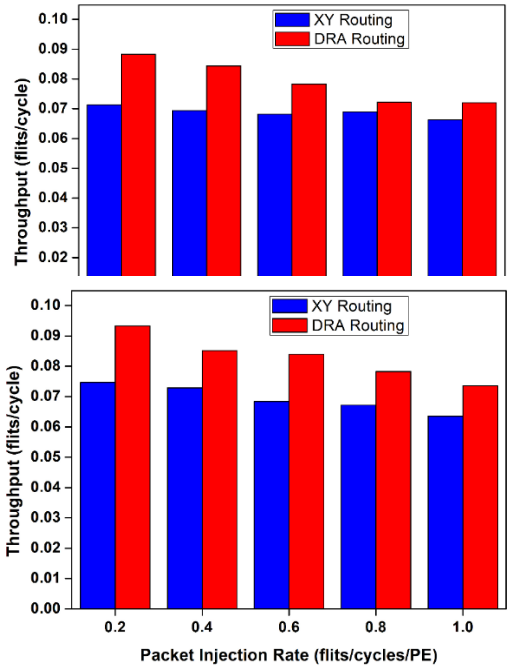

***Fig.10.** Throughput vs. Packet Injection Rate with buffer size=2, 4*

A 16x16 mesh topology throughput and Avg.latency are evaluated for the uniform and random patterns of distinct Buffer Size (BS), using the proposed DRA and conventional XY routing algorithms.Figures.7-10 shows the proposed DRA routing algorithm is performing better than XY routing for Avg.latency and Throughput.

6 CONCLUSION

NoC is now seen as comprehensive solution for addressing the Throughput and Avg.latency problems affecting the current multi-core architectures. This article evaluates the efficiency of the proposed DRA algorithm for 16X16 mesh topology using random and uniform models. By comparing the outcomes, the proposed algorithm is respectively better 18% and 31% for Avg.latency and Throughput.

REFERENCES

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

[2] U. Y. Ogras, P. Bogdan, and R. Marculescu, "An analytical approach for network-on-chip performance analysis," IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, vol. 29, pp. 2001-2013, 2010.

[3] E.-J. Chang, H.-K. Hsin, S.-Y. Lin, and A.-Y. Wu, "Path-congestion-aware adaptive routing with a contention prediction scheme for network-on-chip systems," IEEE Transactions on computer-aided design of Integrated circuits and systems, vol. 33, pp. 113-126, 2013.

[4] S. Foroutan, Y. Thonnart, and F. Petrot, "An iterative computational technique for performance evaluation of networks-on-chip," IEEE Transactions on Computers, vol. 62, pp. 1641-1655, 2012.
[5] K. Yu-Hsin, T. Po-An, H. Hao-Ping, C. En-Jui, H. Hsien-Kai, and W. An-Yeu, "Path-Diversity-Aware Adaptive Routing in Network-on-Chip Systems," pp. 175-182, 2012.
[6] G.-M. Chiu, "The odd-even turn model for adaptive routing," IEEE Transactions on parallel and distributed systems, vol. 11, pp. 729-738, 2000.
[7] W. J. Dally and H. Aoki, "Deadlock-free adaptive routing in multicomputer networks using virtual channels," IEEE transactions on Parallel and Distributed Systems, vol. 4, pp. 466-475, 1993.
[8] P. P. Pande, C. Grecu, M. Jones, A. Ivanov, and R. Saleh, "Performance evaluation and design trade-offs for network-on-chip interconnect architectures," IEEE transactions on Computers, vol. 54, pp. 1025-1040, 2005.
[9] K.-C. Chen, S.-Y. Lin, H.-S. Hung, and A.-Y. A. Wu, "Topology-aware adaptive routing for nonstationary irregular mesh in throttled 3D NoC systems," IEEE transactions on parallel and distributed systems, vol. 24, pp. 2109-2120, 2012.
[10] A. Singh, W. J. Dally, A. K. Gupta, and B. Towles, "GOAL: a load-balanced adaptive routing algorithm for torus networks," in Computer Architecture, 2003. Proceedings. 30th Annual International Symposium on, 2003, pp. 194-205.
[11] L. Schwiebert and R. Bell, "Performance Tuning of Adaptive Wormhole Routing through Selection Function Choice," Journal of Parallel and Distributed Computing, vol. 62, pp. 1121-1141, 2002.
[12] P.-T. Huang and W. Hwang, "An adaptive congestion-aware routing algorithm for mesh network-on-chip platform," in 2009 IEEE International SOC Conference (SOCC), 2009, pp. 375-378.
[13] E. Nilsson, M. Millberg, J. Oberg, and A. Jantsch, "Load distribution with the proximity congestion awareness in a network on chip," in 2003 Design, Automation and Test in Europe Conference and Exhibition, 2003, pp. 1126-1127.
[14] R. Oommen and J. Jose, "Congestion management in adaptive NoC routers using cost-effective selection strategies."
[15] 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.
[16] P. Gratz, B. Grot, and S. W. Keckler, "Regional congestion awareness for load balance in networks-on-chip," in 2008 IEEE 14th International Symposium on High Performance Computer Architecture, 2008, pp. 203-214.
[17] C. Wang, W.-H. Hu, and N. Bagherzadeh, "Congestion-aware Network-on-Chip router architecture," in 2010 15th CSI International Symposium on Computer Architecture and Digital Systems, 2010, pp. 137-144.
[18] J. Kim, D. Park, T. Theocharides, N. Vijaykrishnan, and C. R. Das, "A low latency router supporting adaptivity for on-chip interconnects," in Proceedings. 42nd Design Automation Conference, 2005., 2005, pp. 559-564.
[19] N. Aswathy, R. R. Raj, A. Das, J. Jose, and V. Josna, "Adaptive Packet Throttling Technique for Congestion Management in Mesh NoCs," in International Symposium on VLSI Design and Test, 2017, pp. 337-344.
[20] A. Singh, W. J. Dally, A. K. Gupta, and B. Towles, "GOAL: a load-balanced adaptive routing algorithm for torus networks," in 30th Annual International Symposium on Computer Architecture, 2003. Proceedings., 2003, pp. 194-205.
[21] F. Jafari, Z. Lu, A. Jantsch, and M. H. Yaghmaee, "Buffer optimization in network-on-chip through flow regulation," IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, vol. 29, pp. 1973-1986, 2010.
[22] 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, pp. 1725-1738, 2017.
[23] A. E. Kiasari, Z. Lu, and A. Jantsch, "An analytical latency model for networks-on-chip," IEEE Transactions on Very Large Scale Integration (VLSI) Systems, vol. 21, pp. 113-123, 2012.
[24] R. S. Ramanujam and B. Lin, "Destination-based adaptive routing on 2D mesh networks," in Proceedings of the 6th ACM/IEEE Symposium on Architectures for Networking and Communications Systems, 2010, p. 19.
[25] L. Benini and G. De Micheli, "Networks on chip: A new paradigm for systems on chip design," in Proceedings 2002 Design, Automation and Test in Europe Conference and Exhibition, 2002, pp. 418-419.
[26] S. D. Chawade, M. A. Gaikwad, and R. M. Patrikar, "Review of XY routing algorithm for network-on-chip architecture," International Journal of Computer Applications, vol. 43, pp. 975-8887, 2012.
[27] P. Parandkar, J. Dalal, and S. Kattival, "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, pp. 391-396, 2012.
[28] A. Vitkovskiy, V. Soteriou, and C. Nicopoulos, "A highly robust distributed fault-tolerant routing algorithm for nocs with localized rerouting," in Proceedings of the 2012 Interconnection Network Architecture: On-Chip, Multi-Chip Workshop, 2012, pp. 29-32.
[29] V. Catania, A. Mineo, S. Monteleone, M. Palesi, and D. Patti, "Cycle-Accurate Network on Chip Simulation with Noxim," ACM Transactions on Modeling and Computer Simulation, vol. 27, pp. 1-25, 2016.