A Qualitative Analysis of Various Adaptive Routing Algorithms

Kavyashree G S[1], Saroja S. Bhusare[2], Sunita Shirahatti[3]
Dept. of Electronics and Communication
JSSATE, Bengaluru, India

Abstract—*Network on Chip (NoC) is one of the efficient on-chip communication architecture for System on Chip (SoC) where a large number of computational and storage blocks are integrated on a single chip. NoCs have tackled the disadvantages of SoCs as well as they are scalable. But an efficient routing algorithm can enhance the performance of NoC. In this paper, five different routing algorithms are analyzedsuch as GOAL, RCA, DAR, PDA, and PCA and advantages and disadvantages are compared. All the five algorithms are adaptive routing algorithms.*

Key Words — Multiprocessor system, Network on Chip (NoC), Congestion, Adaptive Routing, Network Topology

1. INTRODUCTION

The interest of humankind is step by step expanding. Individuals dependably incline toward a little electronic gadget having numerous more elements in it. Consequently VLSI industry found another worldview i.e. System on Chip (SoC). As indicated by this worldview diverse electronic or processing frameworks are inserted on a solitary chip. Those processing or electronic frameworks are additionally called as IP cores. Due to the introduction of research into multi-core chips more than a decade ago, on-chip systems have risen as a critical and developing field of examination. As the number of cores are going to increase, there is a relating increment in the bandwidth to encourage high core utilization and scalable on-chip systems. On-chip systems will be predominant in processing areas extending from top of the line servers to installed system on chip (SoC) gadgets.

Many researchers proposed a communication system which is termed as Network on Chip [1] which can avoid problems like scalability, non-adaptive nature, underutilization of resources and less reuse factor. It consists of three important components i.e. Routers, Network Interface (NI) and IP cores or resources. IP cores in the NoC are connected to the network switches. NI (Network Interface) is the communication bridge between the routers and IP cores as routers and IP cores have different communication protocols. For this on chip packet switched network data is converted into some formatted packets and those packets traverse from source to destination with the help of one or more routers in the network. Scalability of this communication system is sufficiently high. It also provides high reusability factor, less complexity and reduced cost. Routing algorithm is an important design concept of Network on Chip. The function of routing algorithm is to determine an efficient path to route the data or packets to transfer from source to destination. Routing algorithms are classified in various basis: a) Deterministic b) Oblivious c) Adaptive routing algorithms. In this paper, ananalysis of five different adaptive routing algorithms are presented. And compared the advantages and disadvantages.

The rest of the paper is organized as follows: Section II presents analysis of various routing algorithms and pros and cons are listed. Section III presents the performance comparison. Section IV concludes the paper.

2. PROPOSALS

Arjun Singh, W J Dally [3] introduced GOAL - Globally Oblivious Adaptive Locally which is adaptive load-balanced routing algorithm for torus networks. This targets on adversarial traffic patterns for achieving higher throughput. On the local traffic, GOAL provides up to 4.6x throughput with locality preservation through randomized routing. For Load Balancing, the route is chosen randomly in all dimensions and with selected directions routing, local load balancing is achieved. GOAL balances the load of network by obliviously choosing the travel direction in every dimension, thereby randomly picking a quadrant for packet transfer. From source to destination, the packet is adaptively routed by travelling only in selected directions. The packet is forwarded in efficient dimension on each hop which has the shortest queue.

Choosing the directions randomly is done using distance-based weights that balances the load in every dimension exactly. GOAL balances the load locally by adaptively routing within that quadrant, once it has been selected. A new algorithm is used by GOAL to get freedom from deadlock using ∗-channels approach extension to handle the non-minimal cases.

GOAL is specifically designed only for TORUS topology and doesn't consider other kind of topologies. The limitation of this method is that the selected direction only uses channel-based information (i.e. output buffer length) to detect path congestion. It doesn't take into account of downstream contention status, and hence cannot decide whether there is sufficient contention at a switch to cause congestion till the buffer is full and blocks the new incoming packets.

Paul Gratz† Boris Grot § Stephen W. Keckler proposed a lightweight technique - "Regional Congestion Awareness (RCA)" to improve network balance globally. Instead of depending only on the information of congestion locally, RCA notifies the routing strategy of congestion in the network away from the neighboring routers. This sustains modest wiring overhead and negligible logic.

RCA which is a methodology that circulates congestion information through the network in ascendable fashion so that it will improve the capability of adaptive routers to spread load of network. The congestion metrics which is locally computed is aggregated with those propagated from neighboring elements before transmitted to upstream routers. This method does not require communication between all nodes and centralized tables that contributes to congestion. Instead, to broadcast congestion information, RCA uses a monitoring network of lower-bandwidth among adjacent routers. This process certainly weighs information of contention by current node distance so that distant congestion impacts routing less than adjacent congestion and reduce the negative impacts of staleness and preventing interference from non-minimal routes.

Although performance improvement are significant with RCA compared to local adaptive routing, this still faces a difficult obstacle of balancing local and remote congestion state. This is because in a minimal adaptive routing, while routing a packet, the congestion on a downstream link heavily depends on whether that link is used to reach the packet's end point. However, in the current routing method it requires additional wire between neighboring routers to observe congestion. In resource-constrained NoCs, Such added overhead may not be tolerable in resource-constrained NoCs.

Rohit Sunkam Ramanujam and Bill Lin proposed a minimal destination-based adaptive routing strategy (DAR). The delay is calculated from each node in the network to every other node, through every candidate output ports and routing choices are made on estimated delay. Regional Congestion Awareness (RCA) [7], previously known as best adaptive routing algorithm which make use of non-local congestion information is outperformed by DAR. Because of per-destination delay estimations in this approach, which are more precise and are not ruined by congestion outside the permissible routing paths to the destination. These delay estimates based on destination provides better control for load-balancing network traffic.

In this approach, using the estimated delay, traffic flows can be controlled independently to various destination on the acceptable paths. Where as in RCA, same congestion metrics were used for the traffic flow going to a common region in the network. Based on their evaluations, it helps to state that RCA is outperformed by DAR in terms of throughput and latency on both synthetic and real workloads.

The working of DAR at a high level is as follows:

- The lone job of router is to distribute the traffic to its routers in next-hop.
- This distribution of traffic is carried out on the basis of per-destination estimate. i.e., the packets using the identical output port and intended for same destination node are distributed to the downstream routers in the same ratio whereas packets using identical output port and intended for different node are distributed independently in different ratio to the downstream routers.
- There exists a set Pj which is having at most two output ports. This set of output port is calculated using estimated delay and is used to distribute the traffic to the downstream router destined for node j. The router will divide the traffic among these two ports in a way that delay to the destination j are equalized through those two ports If set Pj is having only single port, then forcefully all the traffic is routed on to that port.

However, in this routing method, additional wire is required between the adjacent routers to monitor congestion. But in resource-constrained NoCs, additional overhead cannot be acceptable.

Po-An Tsai, Yu-Hsin Kuo[6] proposed adaptive routing scenario with Path-Diversity-Aware (PDA) and Augmented-PDA (A-PDA) selection using path diversity information. To quantify the path diversity characteristics, a formula is derived and number of experiments are conducted with various scenarios. Based on the simulation results, the proposed solution has an edge over other solution in terms of saturation throughput and scalability in large scale Network on Chip. And also router architecture at low cost is proposed for PDA and A-PDA implementation.

During routing, the number of routes that can be taken by packets is represented by Path Diversity (PD) and is affected by network topology and routing algorithms to some extent. By the topology point of view, if the number of connections in network between adjacent nodes are more, then Path Diversity will be higher. In addition, a rectangular mesh is having lower Path Diversity than a square mesh and a short distance source-destination pair will have lower Path Diversity than one of long-distance. And it is said that from the routing algorithm point of view, there exists a higher PD with those having less turn model restriction

Erland Nilsson, Mikael Millberg[7] proposed Proximity Congestion Awareness (PCA) technique. A high performance and low cost switches plays a very important role in the Network on Chip. Theirpaper proposes a memoryless and simple switch for a 2-D regular NoC. Whenever there is a congestion, packets are transmitted in non-ideal direction which is called as deflective routing. PCA is proposed to increase the extreme bearable load of the network, where current switch will make their own switching decisions based on the stressvalue (load information of neighboring switches) thereby avoids the congestion. The current switch will receive four such stress values from its neighboring switches and this helps the current switch to get a clear picture about the surroundings. A situation occurs where, for every other cycle, two neighboring switches will get a high stress value, at that time, the average stress value should be taken over a number of switch of values in order to avoid the oscillations. The simulation result shows that PCA technique will improve the maximum traffic load by a factor of 20 under random traffic.

3. PERFORMANCE COMPARSION

Performance comparison is carried out against these algorithms by considering average load latency and saturation throughput as parameters. RCA shows the router latency reduction of 27% over local and 25% improvement in saturation throughput as estimated by authors [4]. DAR shows 65% of router latency reduction over local and 15% improvement in saturation throughput under Bernoulli injection [5].

GOAL has 40% higher router latency on random traffic then minimal algorithm. GOAL does not shows improvement on saturation throughput on the local patterns [3].

4. CONCLUSION

This paper throws a light on the various adaptive routing algorithms. One thing common in all these routing algorithm is that in order to route the packets these algorithms uses the channel based congestion information in order to decide the path. But it is very important to take concern on the switch congestion and switch contention also. This helps in avoiding the path congestion and improves the quality and effectiveness in the performance of the network.

REFERENCES

[1] C. Duan, V. H. C. Calle, and S. P. Khatri, \Effcient on-chip crosstalk avoidance CODEC design," IEEE Transactions on Very Large Scale Integration (VLSI) Systems, vol. 17, no. 4, pp. 551-560, April 2009.

[2] W. J. Dally and B. Towles, \Route packets, not wires: On-chip interconnection networks," in Proceedings of the Design Automation Conference, Las Vegas, NV, June 18-22 2001, pp. 684-689.

[3] A. Singh, W. J. Dally, A. K. Gupta, and B. Towles, “GOAL: A load-balanced adaptive routing algorithm for torus networks,” in *Proc. Int. Symp. Comput. Arch.*,2003, pp. 194–205.

[4] P. Gratz, B. Grot, and S. W. Keckler, “Regional congestion awareness for load balance in networks-on-chip,” in *Proc. Int. Symp. High-Perform. Comput. Arch.*,pp. 203–214, 2008.

[5] R. S. Ramanujam and B. Lin, "Destination-based adaptive routing on 2D mesh networks," in *Proc. 6th ACM/IEEE Symp. Arch. Netw. Commun. Syst.*, Oct. 2010,pp. 1–12.

[6] Y.-H. Kuo, P.-A. Tsai, H.-P. Ho, E.-J. Chang, H.-K. Hsin, and A.-Y. Wu, "Pathdiversity-aware adaptive routing in network-on-chip systems," in *Proc. IEEE Symp.Embedded Multicore SoCs*, 2012, pp. 175–182.

[7] E. Nilsson, M. Millberg, J. Oberg, and A. Jantsch, "Load distribution with the proximity congestion awareness in a network on chip," in *Proc. Design Autom.Test Eur.*, Mar. 2003, pp. 1126–1127.

[8] *Yuan, R.; Ruan, S.; Gotze, J.,* "*A practical NoC design for parallel DES computation*," in VLSI Design, Automation, and Test (VLSI-DAT), 2013 International Symposium , IEEE Transactions on , vol., no., pp.1-4, 22-24 April 2013

[9] K.-C. Chen, S.-Y. Lin, H.-S. Hung, and A.-Y. Wu, "Topology-aware adaptive routing for non-stationary irregular mesh in throttled 3D NoC systems," *IEEETrans. Parallel Distrib. Syst.*, vol. 24, no. 11, pp. 2109–2120, Oct. 2013.