A REVIEW: LATENCY COMPARISON BETWEEN 2D NOC AND 3D NOC

[1]Neha Jain, [2]Mayank Patel, [3]Sarika Khandelwal

[1]M.Tech Scollarr, [2]Associate professor, [3]Associate Professor
[1] Department of Computer Science & Engineering,
[1]Geetanjali Institute Of Technical Studies, Udaipur, India

Abstract : 3D integration technologies gives new opportunities for Network-on-chip architecture. Customized NoCs are the need of today's SoCs as they offer optimized quality of service and enhanced performances because they are designed to support the specific application behaviour. NoCs designed applications affects the quality metrics and overall communication latency of the system.It is a heuristic based on Branch-and-Bound approach. It is used for latency aware smart application to core mapping in 3D Mesh Network-on-chip. The proposed methodology reduced the average latency consumed in the optimally mapped 3D Mesh in comparison to optimally mapped 2D Mesh of same size.

IndexTerms **– 3D NOC, Latency, Heuristic Mapping**

I. INTRODUCTION

The upcoming generation SoCs will comprise of a huge number of cores and the key challenge to work out will be the NoC bottleneck of these systems which confines scalability. The onset of 3D integration technologies has unlocked the doors of novel prospects for design of on-chip networks in SoCs. The union of two evolving standards, NoC & 3D IC, enables the design of novel design structures with significant performance enhancements in quality metrics upon conventional solutions [8]. To attain this, in this paper we have formulated mapping problem followed by demonstration of the result of several applications to cores mappings on the dynamic communication latency of a given system. In this paper a Branch-and-Bound heuristic to intelligently map the given set of application onto cores in 3D NoC architecture to reduce the average dynamic communication latency of the system is presented. In order to validate efficiency of the proposed approach several experiments have been carried out on several arbitrary scales. In [4][10] it is illustrated that in addition to the footprint reduction in a fabricated design, 3D network structures are more inclined towards leading enhanced performance in terms of smaller latency, lower dissipation of energy and higher throughput in comparison to conventional 2D NoC archetypes. In [6] the mapping problem for 2D regular Tile - based structural designs is addressed.

II. LATENCY

The proposed approach makes use of the model presented in [6]. The design of chip is comprised of PxQxR tiles which are linked as per the fundamental 3D Mesh structure. Each tile in 3D NoC has IP Core, Virtual Channels (VCs) & seven communications links (East, West, North, South, Front, Rear and Core). The model proposed in [6] takes the Manhattan distance between the cores into consideration while mapping the applications. The basic idea is to map the cores that communicate with each other at the smallest Manhattan distance as possible. Moreover as the NoC architecture is a communication centric design, both these factors ultimately lead to the reduced average latency of the system and energy as well in comparison to a randomly mapped 3D Mesh NoC.

In order to minimize the overall latency of system, we need to obtain a one to one mapping of applications onto the cores in 3D Mesh NoC.

Core Graph, CG, G = G(C, Rp) is a directed graph consisting of set of vertices C, where c_i represents a core in the architecture, and $Rp_{i,j}$ represents a directed edge that is the routing path computed using XYZ routing algorithm between $c_i$and c_j. e($Rp_{i,j}$) refers to the consumption of average energy (joule) in carrying a bit of data from core c_i to c_j. Set of links that constitute the $Rp_{i,j}$ is represented by L ($Rp_{i,j}$).

Application Graph, AG, G = G(A, T) is a directed graph consisting of set of vertices A where ap_i refers to an application and each directed edge $t_{i,j}$representsthat a_i communicates with a_j. Communication volume (bits) between any two applications is represented by V($t_{i,j}$). The least possible bandwidth (bits/sec.) which the underlying communication structural design should offer is denoted by bw ($t_{i,j}$).

The heuristic we have proposed in this paper is built on the concept of branch and bound approach. The heuristic travels through a search tree that represents the solution space with the aim of obtaining an optimal mapping with the least communication cost. This could be only possible if the two applications communicating with each other can be placed as near as possible. If two cores are placed nearly then the time taken by the data to reach from source to destination will also reduce. A label is assigned to every node in the tree. For example, node 356xxxxx implies an internal node where Core3, Core5 and Core6 of CG are mapped with application number A0, A1 and A2 of AG in that order; and unmapped applications are A3 to A7. A data-traffic matrix is maintained that stores the communication requirements which comprise incoming plus outgoing data traffic from a particular application to every other application in given AGThe mapping cost of the child nodes are always greater than that of parent nodes and based on this, unqualified tree-branches are clipped later on. A node is legal if it encounters the requirement in terms of bandwidth amongst the mapped applications [6][9]. Illegal parent node produces illegal child nodes.

The upper end cost (UEC) of a node stands for a cost that is not lower than the minimum communication cost of its legitimate successor child nodes comprising of all the applications that have to be mapped (i.e. leaf nodes). So as to calculate as lowest UEC as achievable for a node, a greedy methodology for the applications to cores mapping is implemented.

The lower end cost (LEC) of a node stands for the best achievable communication cost that its legal descendant child nodes comprising of all the applications that have to be mapped (i.e. leaf nodes) can probably reach.

The heuristic goes over the following two stages until it attains an optimal solution.

Branch: Under this step, the next unexpanded node from the front of node preference queue is removed and explored further to generate new child nodes by mapping the next application that has not been mapped yet and has the extreme communication requirement; onto a core from the set of vacant cores.

Bound: Under this step, each child node that has been produced in the aforementioned step is examined to comprehend if it leans towards yielding the best leaf nodes later. For this node the UEC and LEC costs are evaluated and this node is clipped if either the communication cost among applications that have already been mapped on cores or results in higher LEC than the lowest UEC that has been estimated all through the exploration.

III. EXPERIMENTAL RESULTAND ANALYSIS

NC-G-SIM simulator is used in order to test the performance of the heuristic presented in this paper. NC-G-SIM is a discrete event, cycle accurate simulator which supports Regular 3D, 2D and irregular topology framework with XYZ and distributed table based routing. During experimenting the performance on the 3D Mesh NoC, the number of maximum slices kept is 4 for the reason that in 3D ICs as the number of vertically stacked dies grows, power density/area and the length of heat conduction path also rises [1][3][5].Estimating using Orion [7] for 0.18µm technology C_{LB} is set to 0.0007 andC_{SB} is assumed to be 0.54 and 0.52 for 6 Ports and 4Ports router respectively. The packet size is 8 bytes and flit-interval is set to 2 clock cycles. The heuristic is tested on different topology sizes where number of cores used ranges from 8 to 512. The heuristic is implemented on 5 sets of 100 varying topologies each to get the intelligent along with random application to cores mapping in both 3D and 2D Mesh NoC structural designs of similar sizes with similar traffic settings. With the help of TGFF [2] thousand sets of benchmarks were arbitrarily generated with varied requirement in terms of bandwidth and communication volume of the IP cores in line with the specified distribution. The routing schemes used are XY and XYZ to route the traffic in 2D and 3D Mesh NoC respectively. The average latency/flit is taken as performance metric.

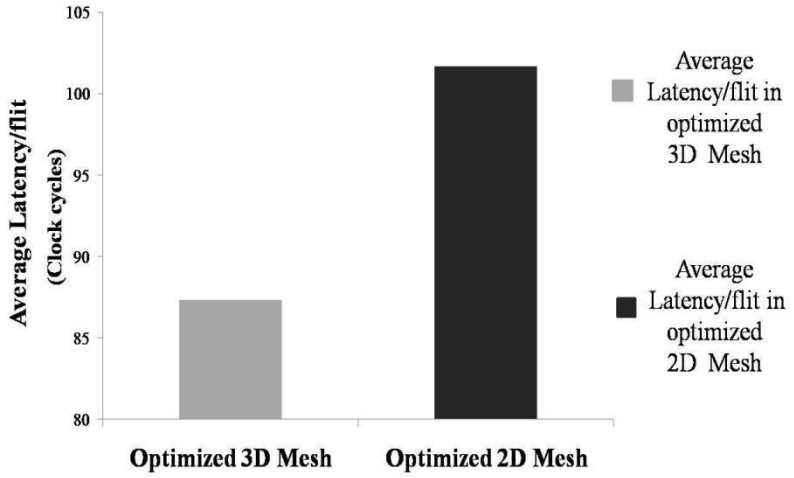

Figure1: Comparison of Latency between optimized mapping same sized 3D and 2D Mesh topologies with similar traffic conditions

The graph plotted in in fig.1 clearly shows that average latency per flit has reduced by a reasonable amount in the intelligently mapped 3D Mesh NoC in comparison to intelligently mapped 2D Mesh NoC as well.

In comparison to 2D Mesh NoC, the proposed heuristic results in 14.9-33.9% savings in average delay in the reaching the packets at their destination(i.e. average latency/flit).

IV. CONCLUSION:

This review paper presents an intelligent latency aware mapping heuristic to map the applications in the given AG onto the cores in the CG in 3D Mesh NoC archetype. The proposed methodology reduced the average latency consumed in the optimally mapped 3D Mesh in comparison to optimally mapped 2D Mesh of same size.

V. REFERENCES:

[1] Bernstein, K., Andry, P., Cann, J., Emma, P., Greenberg, D., Haensch, W., & Young, A. (2007, June). Interconnects in the third dimension: Design challenges for 3D ICs.In Proceedings of the 44th annual Design Automation Conference (pp. 562567).ACM.Choudhary, N., Gaur, M. S., &Laxmi, V. (2011).Energy Efficient Network Generation for Application Specific NoC. Global Journal of Computer Science and Technology, 11 (16).

[2] Dick, R. P., Rhodes, D. L., & Wolf, W. (1998, March). TGFF: task graphs for free. In Proceedings of the 6th international workshop on Hardware/software codesign (pp. 97-101). IEEE Computer Society.

[3] Ebrahimi, M., Daneshtalab, M., Liljeberg, P., Plosila, J., &Tenhunen, H. (2011, May). Exploring partitioning methods for 3D Networks-on-Chip utilizing adaptive routing model.In Networks on Chip (NoCS), 2011 Fifth IEEE/ACM International Symposium on (pp. 7380).IEEE.

[4] Feero, B. S., &Pande, P. P. (2009). Networks-onchip in a three-dimensional environment: A performance evaluation. Computers, IEEE Transactions on, 58 (1), 32-45.

[5] Hassanpour, N., Khadem, P., Hessabi,S. (2013). A Task Migration Technique for Temperature Control in 3D NoCs. In 27th IEEE International Conference on Advanced Information Networking and Applications (AINA). Manuscript submitted for publication.

[6] Hu, J., Marculescu, R. (2005). Design methodologies for application specific networks-on-chip. Ph.D. dissertation.Carnegie Mellon University.

[7]. Kahng, A. B., Li, B., Peh, L. S., &Samadi, K. (2009, April). Orion 2.0: A fast and accurate noc power andarea model for early-stage design space exploration. In Proceedings of the conference on Design, Automation and Test in Europe (pp. 423-428). European Design and Automation Association.

[8] Rahmani, A. M., Latif, K., Liljeberg, P., Plosila, J., &Tenhunen, H. (2010, November). Research and practices on 3D networks-on-chip architectures. In NORCHIP, 2010 (pp. 1-6). IEEE.

[9] Wadhwani, P., Choudhary, N. and Singh D. (2013). Energy Efficient Mapping in 3D Mesh Communication Architecture for NoC.International journal of Global Journal of Computer Science and Technology on Network, Web & Security (pp. 1-6).

[10] Xu, Y., Du, Y., Zhao, B., Zhou, X., Zhang, Y., & Yang, J. (2009, February). A low-radix and low-diameter 3D interconnection network design.In High Performance Computer Architecture, 2009.HPCA 2009. IEEE 15th International Symposium on (pp. 30-42). IEEE.