Routing Centric NoC Design for High Performance Multimedia Application

Naveen Choudhary, M. S. Gaur, V. Laxmi

Abstract*— NoC has been proposed as a solution for the communication challenges in the nanoscale regime of SoC. In order to tackle design complexity and to facilitate reuse, systems are typically required to be built from pre-designed and pre-verified homogenous or heterogeneous building blocks such as programmable RISC cores, DSPs, memory blocks. Most SoC platforms are special-purpose tailored to the domain-specific requirements of their application, which communicate in a very specific, mostly irregular way. In this work, we propose a methodology for routing centric Network-on-Chip design for High Performance Multimedia Application. The proposed methodology exploits a priori knowledge of the applications communication characteristic to generate an optimized network topology along with chosen routing function compliant routing tables to improve communication performance by improved traffic load distribution.

Index Terms*— NoC, SoC, ULSI, on-chip networks, application optimized NoC.

I. INTRODUCTION

Due to the continued shrinking of feature sizes in today's silicon technologies, the integration of complex systems-on-chip (SoC) has become feasible today, offering a tremendous amount of computational power. In order to tackle design complexity and facilitate reuse, these systems are typically built from pre-designed and pre-verified heterogeneous building blocks like programmable RISC cores, DSPs, dedicated hardware accelerators, memory blocks etc. which are plugged together in a domain specific integration platform. These designs typically show a high degree of modularity and inherent computational parallelism. This trend is accompanied by revolutionary changes of the employed design methodology where a paradigm shift from a computation-centric view to a communication-centric becomes evident [1], [2]. Functionality of such systems is often captured by a set of communicating tasks at a high level of abstraction. These are mapped to computational resources which are then interconnected by a central communication backbone. The efficiency of the overall design is governed by this communication architecture which plays a key role in modern SoC platforms. It impacts both performance and implementation costs in terms of silicon area and energy consumption to a substantial extent.

Manuscript Received September 25, 2011

Naveen Choudhary, Department of Computer Science and Engineering, College of Technology and Engineering, Maharana Pratap University of Agriculture and Technology, Udaipur, Rajasthan, India, Mobile No: +91-9352356102, (e-mail: naveenc121@yahoo.com).

M. S. Gaur, Department of Computer Engineering, Malaviya National Institute of Technology, Jaipur, India, (e-mail: gaurms@gmail.com).

V. Laxmi, Department of Computer Engineering, Malaviya National Institute of Technology, Jaipur, India, (e-mail: vlaxmi@mnit.ac.in).

Some of the most important phases in designing the NoC are the design of the topology or structure of the network and setting of various design parameters (such as frequency of operation, link-width, etc). Several early works [1], [2] favored the use of standard topologies such as meshes tori, and under the assumption that the wires can be well structured in such topologies These approaches are adequate for general purpose systems where the traffic characteristics of the system cannot be predicted statically, as in homogeneous chip-multiprocessors [8]. However, most SoCs are heterogeneous, with each core having different size, functionality and communication requirements. Thus, standard topologies can have a structure that poorly matches application traffic. This leads to large wiring complexity after floorplanning, as well as significant power and area overhead. Moreover, for most SoCs the system is designed with static (or semi-static) mapping of tasks to processors and hardware cores and hence the communication traffic characteristics of the SoC are well characterized at design time. In addition to above mentioned applicability of irregular topologies in application specific NoCs, the generic regular topologies can also become irregular for supporting oversized region or due to faults in switches or links in the regular NoCs. The region concept presented in [17] was intended for use of larger resources which do not fit in the fixed sized slot of a regular mesh architecture layout as shown in figure 3. Nevertheless, this concept can be used in a variety of other contexts [18] as mentioned below.

- Encapsulating a group of resources with special requirements on performance, power consumption or data security. Such a region could have specialized interconnections as well as communication protocols.
- Region as a logical structure. In this case the internal hardware design of the region is identical with the outside NoC structure. This assumes configurable routers in the NoC for defining, isolating and maintaining a region.
- Support for different configurations of power/ performance modes of resources inside a region by control of operating voltage, clock frequency etc.
- Reuse of multi-core subsystems. These solutions are currently available as separate SoCs. The concept of region offers the possibility of raising the level of reuse from a core to a level where specially designed multi-core subsystems can be reused. Without the region concept these subsystems will need to be redesigned, keeping in view the NoC constraints.

If the topology is regular, it is wise with regards to performance, to use a topology dependent routing since it would be able to exploit the regularity of the topology. Unfortunately, these algorithms are sensitive to topology changes. A faulty switch or link will degrade the topology into an irregular one and then the algorithms will fail. A simple way to achieve fault-tolerance is by the use of a topology agnostic routing algorithm probably in the combination with static reconfiguration if required.

Deadlock free communication is a significant requirement for a good NoC. There are many popular turn prohibition [3] based deadlock free topology agnostic routing algorithms such as up*/down* [4], Left-Right [5], L-turn[5]

Application-specific custom topology design has been explored in [6]-[8], [9], [10]. The works from [6], [7] do not consider the floorplanning information during the topology design process. In [10], a floorplanner is used during topology design to reduce power consumption on wires. However it does not consider the area and power consumption of switches in the design. Also, the number and size of network partitions are manually fed. In [8], a slicing tree based floorplanner is used during the topology design process. This work assumes that the switches are located at the corners of the cores and it does not consider the network components (switches, network interfaces) during the floorplanning process. Moreover the actual sizes of the cores in [8], [9] are considered only after generating their relative positions. The resulting floorplan can be extremely area inefficient when compared to the standard floorplanning process.

In the proposed work, a routing function based irregular topology generation methodology is developed for customized design irregular communication infrastructure for high performance SoC according to the communication requirement of the application for optimized run time performance of the SoC. Irregular NoC communication model and architecture is defined in Section II. The proposed routing function centric topology design methodology is presented in Section III. Section IV summarizes experimental results followed by a brief conclusion in section V.

II. COMMUNICATION MODEL AND ARCHITECTURE

The generic communication model including Task graphs [11], [12], Core Graph, and NoC topology/interconnection network is shown in Figure 1 Where *Core Graph* is a directed graph with vertices representing the IP cores and the directed edges between vertices represents the communication bandwidth requirement of the application. Similarly *NoC topology graph* is a directed graph with each vertex representing a core/node in the topology and a directed edge among vertices represents the direct available communication channel along with the actual physically available bandwidth in the topology.

For Performance evaluation, a discrete event, cycle accurate simulator *IrNIRGAM* can be used for Irregular NoC. *IrNIRGAM* is an extension of *NIRGAM* [13]. Ir*NIRGAM* is a cycle-accurate SystemC based performance simulator which supports irregular topology framework with table based routing.

For heterogenous NoC, the chip layout can be decided with the help of floorplanning according to desired metric such as area as a pre-processing step. The energy model given in [11] can be extended for unequal (varying) channel length for Irregular Network-on-Chip as shown below.

$$E_{bit}(t_i, t_j) = n_{hops} \times Er_{bit} + \sum_{k=1}^{n_{hops}-1} El_{bit}^{k} \quad ----(1)$$

Where $E_{bit}(t_i, t_j)$ is the average dynamic energy consumption for sending one bit of data from tile t_i to tile t_j, n_{hops} is the number of routers the bit traverses from tile t_i to tile t_j, Er_{bit} is the energy consumed by router for transporting one bit of data and El_{bit} is the energy consumed by each link/channel in the route, the bit follows from communication *source core* to the *destination core*.

In this paper deterministic version of up*/down* [5] routing function is chosen for optimized design of the routing function centric topology design.

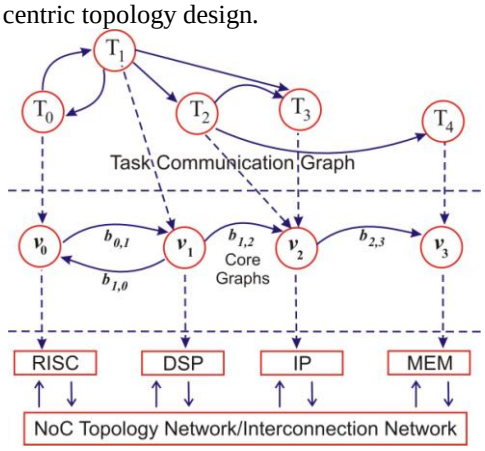

Fig. 1. Communication model for application customized Network-on-Chip

III. ROUTING FUNCTION CENTRIC TOPOLOGY DESIGN METHODOLOGY

Based on the information from chiplayout and traffic characteristics of the application, the global and detailed physical routes for the customized NoC are generated using the proposed methodology assuming over the cell routing [14]. Irregular topology construction is initiated by creating a minimum spanning tree (*MST*) based on Manhattan distance among the IP cores with root as the node/core having maximum communication requirement. Moreover the permitted node degree (nd_tree_{max}), i.e., number of allowed ports per IP core in initial stage of the methodology is kept less than the actual permitted node degree (nd_{max}) to allow better search space for valid shortcuts. The *MST* helps in classifying all the channels of the topology as "*up*" or "down" in addition to making the initial topology strongly connected, providing a path between every pair of nodes. In the next phase of the methodology a genetic algorithm [15] based heuristic is used for the extended design of customized NoC/topology.

Genetic algorithm [15] is a search technique used in determining exact or approximate solutions to optimization and search problems.

The generated customized NoC topology is expected to exhibit reduced congestion and average flit latency leading to increased throughput for the application specific injected traffic. In the proposed methodology the link/channel length is not allowed to exceed the maximum permitted channel length (e_{max}) due to constraint of physical signaling delay. The nodes of the generated topology are not allowed to exceed a given maximum permitted node-degree (nd_{max}). This constraint prevents the algorithm from instantiating slow routers with a large number of I/O-channels which would decrease the achievable clock frequency due to internal routing and scheduling delay of the router. Figure 2 briefly illustrates the proposed methodology.

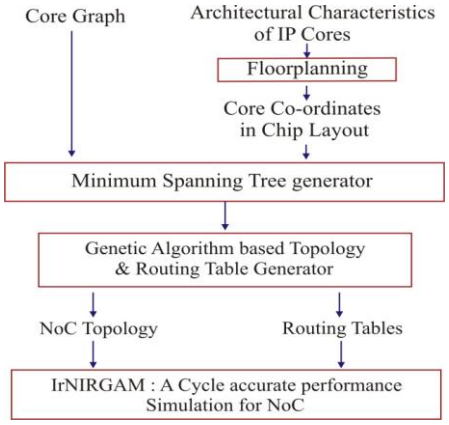

Fig. 2. NoC construction flow using genetic algorithm

A. Initial Population Generation and Solution Representation

Modified dijkstra's shortest path algorithm is used to find energy shortest deadlock free path in accordance to the up/down (*Left-Right*) rule in the NoC topology graph (*MST* in the initial topology). Routing table entries for the routers of the NoC is generated for each *traffic characteristic* (edge) in the *Core Graph*. At this stage the traffic load to these tree paths is assigned according to the bandwidth requirement of the *traffic characteristics*. Moreover to bring variety as well as to include the possible shortest deadlock free paths in the topologies of the initial population, a large number of genes of initial population are mutated by laying energy shortest deadlock free path in accordance to the *up*/down** rule with constraints nd_{max} and e_{max} firmly kept.

In the proposed formulation, each chromosome is represented by an array of genes. Maximum size of the gene array to be equal to the number of *traffic characteristics* (i.e. edges) in the *Core Graph*, in other words a chromosome represents an instance of NoC topology and each gene represents a collection of deadlock free paths with upper limit of *n* (configurable parameter) for a *traffic characteristic* in the *Core Graph* along with necessary information for these paths. In each gene at least one path is the shortest energy path through the channels exclusively pertaining to *MST*, guarantying the connectivity between the source and destination pair of the gene (*traffic characteristics*).

B. Genetic Operations

Three mutation operations called *Topology Extension* (i.e. topology is extended by laying additional paths for a randomly selected gene), *Topology Reduction* (i.e. topology is pruned by removing paths of the topology which are very lightly loaded with traffic) and *Energy Reduction* (i.e. a random number of selected paths of the topology are replaced with energy conscious shortest paths) with equal probability are applied in each generation of the genetic algorithm.

In addition to the above the crossover operation is performed on a large size of the population with the bias towards the *Best Class* of the chromosome population. For achieving crossover of two chromosomes, a random crossover point is selected and then genes of these chromosomes are mixed over the crossover point to produce two new chromosomes.

C. Fitness Function

The fitness measure essentially has two components - (1) *average bandwidth requirement overflow* in comparison to preferred bandwidth load, (2) *dynamic communication energy requirement* of the traffic for the customized NoC. The fitness (cost) function can be formulated as under.

$$Cost_i = \alpha \times (Ec_i / X_1) + (1-\alpha) \times (Bc_i / X_2)$$

Where X_1 be maximum chromosome dynamic communication energy requirement, X_2 be maximum possible bandwidth requirement of a channel, Ec_i is the energy requirement, Bc_i is the average bandwidth requirement overflow per channel for chromosome c_i and α is an empirical constants.

Through exhaustive experimentation, the optimum value of α was determined as 0.4. Fitness of chromosome is regarded as high if its cost approaches zero. The chromosome with the *best Cost* is selected as the output chromosome of the genetic algorithm based proposed methodology.

IV. EXPERIMENTAL RESULTS

The generated routing centric customized topology was evaluated on *IrNIRGAM* simulation framework for a realistic multimedia application as the benchmark. However TGFF [12] if required can be used to generate random test bench marks. The proposed Routing Centric Custom NoC design (*RC-NoC*) was run for 1000 generation with population size of 500. The mutations are done on 45% of the population and crossover on 35% of the population in each generation. For performance comparison, the NoC simulator *IrNIRGAM* was run for 10000 clock cycles and network throughput in flits, average flit latency, traffic load and energy distributions per channel were used as parameters for comparison. Further traffic load per channel and energy per channel exhibits the traffic load in flits per channel and communication energy consumed per channel respectively for the simulation run.

The dynamic communication energy consumption by router in transmitting a bit is evaluated using the power simulator orion [16] for 0.18µm technology. Moreover the dynamic bit energy consumption for inter-node links (El_{bit}) can be calculated using the following equation.

$$El_{bit} = (1/2) \times \alpha \times C_{phy} \times V_{DD}^2$$

Where α is the average probability of a 1 to 0 or 0 to 1 transition between two successive samples in the stream for a specific bit. The value of α can be taken as 0.5 assuming data stream to be purely random. C_{phy} is the physical capacitance of inter-node wire under consideration for the given technology and V_{DD} is the supply voltage.

The proposed *RC-NoC* was compared with regular 2D-Mesh NoC for realistic multimedia application (*MMS*) For *RC-NoC*, the *permitted channel length* (e_{max}) was taken as 2 times the length of the core/node and *permitted node/core degree* (nd_{max}) of 4 was assumed with *up*/down** routing function. *MMS* is an integrated video/audio system which includes an h263 video encoder, an h263 video decoder, an mp3 audio encoder and an mp3 audio decoder. The application was partioned into 40 distinct tasks and then these tasks were assigned and scheduled onto 25 selected IPs. These IPs range from DSPs, generic processors, embedded DRAMs to customized ASICs. The communication trace graphs as shown in Figure 3 for the same were obtained from the work presented by Hu et al. [11].

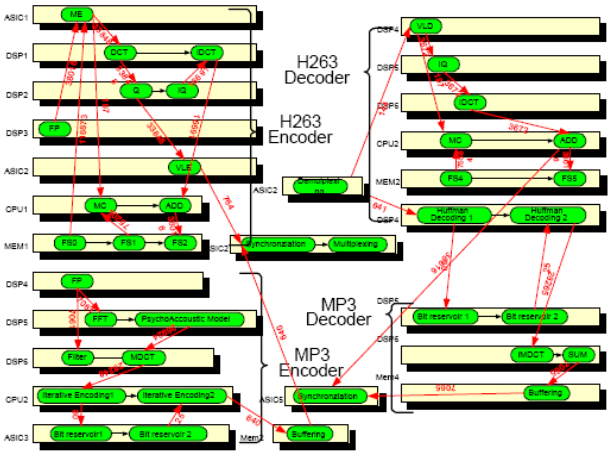

Fig. 3. MMS Communication Trace Graph

Figure 4 shows that *RC-NoC* with *up*/down** routing function exhibit an increase in throughput of 30.5% and reduction in average flit latency of 6% and 22.8% in comparison to *2D-Mesh* with XY and OE routing respectively for *MMS* application. Moreover the *RC-NoC* exhibits better distribution of application specific traffic across the channels of the generated topology as is evident from Figure 5(a, b).

Figure 5(a, b) exhibits the effect of traffic load and energy distribution across the channels of the network for 1000 injected flits into the NoC according to the application requirement. RC-NoC with up*/down* routing function shows reduction in average traffic load per channel of 22% and 36.8% and reduction in average per unit length communication energy consumption per channel of 21.2% and 36.6% in comparison to 2D-Mesh with XY and OE routing respectively for the realistic *MMS* application.

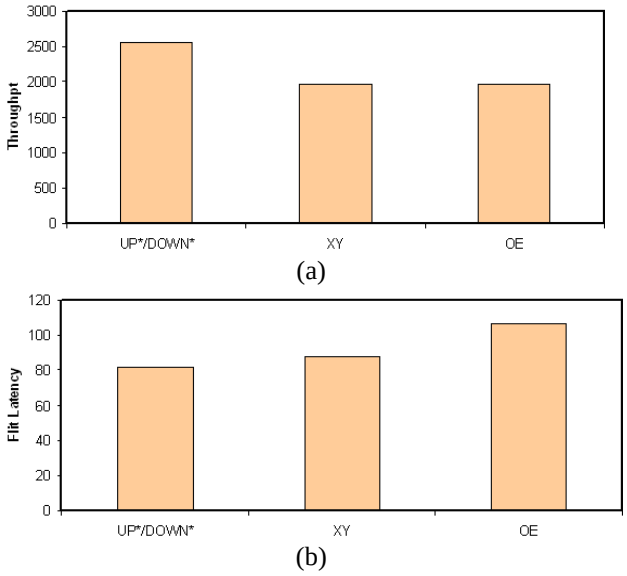

Fig. 4. Performance results of *RC-NoC* (*up*/down) and regular *2D-Mesh* on realistic *MMS benchmark* for (a) throughput (in flits) and (b) flit latency (in clocks)**

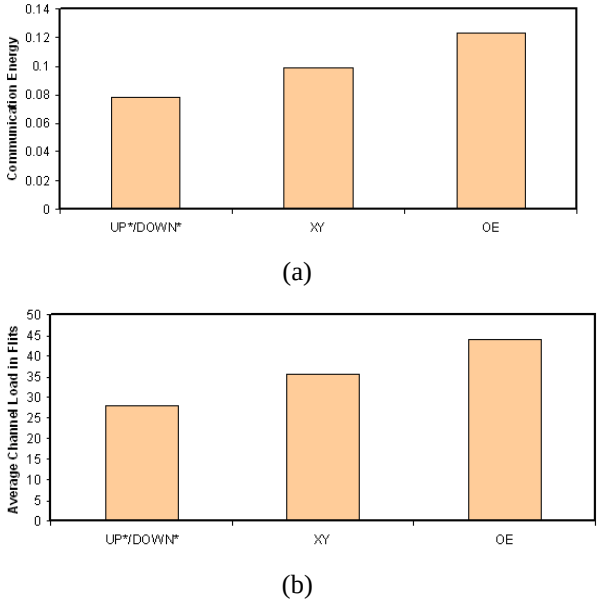

Fig. 5. Comparison of traffic distribution across the channels of *RC-NoC* (*up*/down) and regular *2D-mesh* on realistic MMS benchmark (a) average per channel communication energy consumption (in pico joules) (b) average per channel communication traffic load (in flits)**

V. CONCLUSION

In the presented work, routing centric customized NoC design methodology based on application communication requirement is proposed. The presented methodology is adaptable according to any routing function where generic routing rules can be enforced for the routing function.

The methodology was tested on realistic multimedia benchmark and quite encouraging results were achieved. It is believed that if other NoC design parameters such as number of virtual channels, buffer size, chosen switching scheme are considered at the time of designing a customized topology according to the application requirement can result in huge performance gain.

REFERENCES

1. W. J. Dally and B. Towles, "Route Packets, Not Wires: On-Chip Interconnection Networks," *IEEE Proceedings of 38th Design Automation Conference (DAC)*, 2001, pp. 684–689.
2. S. Kumar, A. Jantsch, J. P. Soininen, M. Forsell, M. Millberg, J. Oberg, K. Tiensyrja, and A. Hemani, "A Network on Chip Architecture and Design Methodology," *Proceedings of Very Large Scale Integration (VLSI) Annual Symposium* (*ISVLSI 2002*), 2002, pp. 105–112.
3. C. Glass and L. Ni, "The Turn Model for Adaptive Routing," *Proceedings of 19th International Symposium on Computer Architecture*, May 1992, pp. 278– 287.
4. M. D. Schroeder et al. "Autonet: A High-Speed Self-Configuring Local Area Network Using Point-to-Point Links," *Journal of Selected Areas in Communications*, vol. 9, October 1991.
5. A. Jouraku, A. Funahashi, H. Amano and M. Koibuchi, "L-turn routing: An Adaptive Routing in Irregular Networks," *Proceedings of International Conference on Parallel Processing*, September 2001, pp. 374-383.
6. Pinto et al. "Efficient Synthesis of Networks on Chip". *Proceedings of ICCD*, October 2003, pp. 146-150.
7. W. H. Ho and T. M. Pinkston, "A Methodology for Designing Efficient On-Chip Interconnects on Well-Behaved Communication Patterns," *Proceedings of HPCA*, February 2003, pp. 377-388.
8. K. Srinivasan et al. "An Automated Technique for Topology and Route Generation of Application Specific On-Chip Interconnection Networks," *Proceedings of ICCAD 2005*, 2005.
9. K. Srinivasan and K. S. Chatha, "ISIS: A Genetic Algorithm based Technique for Custom On-Chip Interconnection Network Synthesis," *Proceedings of 18th International Conference on Very Large Scale Integration (VLSI) Design*, Kolkata, India, 2005, pp. 623-628.
10. T. Ahonen et al. "Topology Optimization for Application Specific Networks on Chip," *Proceedings of SLIP 2004*, 2004.
11. J. Hu, and R. Marculescu, "Energy-Aware Mapping for Tile-based NOC Architectures under Performance Constraints," *proceedings of ASP-DAC* 2003, Jan 2003.
12. R. P. Dick, D. L. Rhodes and W. Wolf, "TGFF: Task Graphs for Free," *Proceedings of International Workshop on Hardware/Software Codesign*, March 1998.
13. Lavina Jain, B. M. Al-Hashimi, M. S. Gaur, V. Laxmi and A. Narayanan, "NIRGAM: A Simulator for NoC Interconnect Routing and Application Modelling, *Proceedings of DATE 2007*, 2007.
14. K. Srinivasan and K. S. Chatha, "Layout Aware Design of Mesh based NoC Architectures," *Proc. of 4th International Conference on Hardware Software Codesign and System Synthesis*, Seoul, Korea, 2006, pp. 136-141.
15. A. E. Eiben and J. E. Smith, *Introduction to Evolutionary Computing*, Springer-Verlag, Berlin, Heidelberg, 2003.
16. A. B. Kahng, B. Li, L. S. Peh and K. Samadi, "Orion 2.0: A Fast and Accurate NoC Power and Area Model for Early-Stage Design Space Exploration," *Proceedings of DATE'09*, 2009, pp. 423–428.
17. S. Kumar, A. Jantsch, J.P. Soininen, M. Forsell, M. Millberg, J. Öberg, K. Tiensyrjä, A. Hemani, "A Network on Chip Architecture and Design Methodology", In IEEE Annual Symposium on VLSI, April 2002.
18. R. Holsmark, S. Kumar, "Design Issues and Performance Evaluation of Mesh NoC with Regions", In IEEE NorChip, Oulu, Finland, pp. 40-43, Nov. 2005.

AUTHORS PROFILE

Dr. Naveen Choudhary received his B.E, M.Tech and PhD degree in Computer Science & Engineering. He completed his M.Tech from Indian Institute of Technology, guwahati, India and PhD from Malviya National Institute of technology, Jaipur, India in 2002 and 2011 respectively. Currently he is working as Associate Professor and Head, department of Computer Science and Engineering, College of Technology and Engineering, Maharana Pratap University of Agriculture and Technology, Udaipur, India. His research interest includes Interconnection Networks, Network on Chip, Distributed System and Information Security. He is a life member The Indian Society of Technical Education, Computer Society of India and The Institution of Engineers, India. E-mail: naveenc121@yahoo.com

Dr. M. S. Gaur received his B.E. from MBM Engineering College, Jodhpur, India. He completed his M. E. from IISC, Bangalore, India and PhD from University of Southampton, UK. Currently he is working as Professor in the Department of Computer Engineering, MNIT, Jaipur. He is also Dean, Student welfare at MNIT, Jaipur, India. His research interest includes Networks on Chip, Network Simulation and Information Security. E-Mail: gaurms@gmail.com

Dr. V. Laxmi received her B.E. from MBM Engineering College, Jodhpur, India. She completed her M. Tech from IIT, Delhi, India and PhD from University of Southampton, UK. Currently she is working as Reader in the Department of Computer Engineering, MNIT, Jaipur. She is also Head, Department of Computer Engineering, MNIT, Jaipur, India. Her research interest includes Algorithms, Networks on Chip, Image processing and Information Security. E-Mail: vlaxmi@mnit.ac.in