Migration of On-Chip Networks from 2 Dimensional Plane to 3 Dimensional Plane

Naveen Choudhary

Abstract*—In spite of *the higher scalability and parallelism presented by 2D-Network-on-Chip (NoC) over the conventional shared-bus based systems, it is still not an ultimate solution for future large scale Systems-on-Chip (SoCs). Recently, NoC integration in three dimensions is (3D-Network-on-Chip) proposed as a potential solution offering higher speed, low latency, lower dynamic power consumption and high parallelism Advanced integration technologies are making feasible the extension of topology synthesis of on-chip networks from 2 dimension to 3 dimension. Studies have highlighted that 3D NoCs can significantly improve communication efficiency due to reduced communication distances in 3D space.

This paper presents a brief journey of research in the domain of Network on Chip topology synthesis from 2D dimensional plane to 3 dimensional plane and highlights the major challenges and issues faced and addressed by the NoC research community in the design of 2D standard NoCs, irregular & application specific 2D NoCs and 3D NoCs.

***IndexTerms*— Application-Specific-NoC, 3D-NoC, 2D-NoC, On-Chip networks, System-on-Chip.**

I. INTRODUCTION

The fast emerging integrated circuit fabrication technology has made possible fabrication of billions of transistors on a single chip. The communication complexities of system on chip is increasing exponentially as the number of transistors on a chip is increasing. In such a senerio, the conventional bus-based System-on-Chip (SoC) will not be effective due to communication performance bottlenecks. Most SoCs used to have bus-based interconnection architectures, such as simple, hierarchical or crossbar-type buses. Bus based systems do not scale well with the system size in terms of bandwidth, clock frequency and power consumption [1]. To address such on-chip communication issues the Network on Chip was proposed as a promising solution for future on chip multicore systems. Macro-network communication practices were inherited for on-chip communication. Fig. 1 exhibits a 4×4 mesh based Network on Chip interconnection network. The network is comprised of network links providing raw bandwidth and routers (R) providing logic for choosing appropriate paths during communication, each router is connected to an IP/core/ processing element/modules with the help of a network interface (NI) logic. The communication among cores is achieved with the help of packet transmission over the underlying topology/interconnection network.

Manuscript received on April, 2013.

Naveen Choudhary, Department of Computer Science and Engineering, College of Technology and Engineering, Maharana Pratap University of Agriculture and Technology, Udaipur, Rajasthan, India.

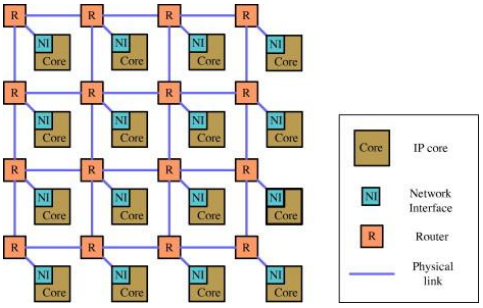

Fig. 1: 4×4 Network on Chip

Early research in the domain of NoC prefered the use of standard topologies such as meshes, tori, k-ary n-cubes or fat trees under the assumption that the wires can be well structured in such NoCs. Moreover the regular NoC also supported the design and fabrication of generic components, which can be easily mass produced by the industry. Standard and regular NoC were fine for general purpose systems with homogenous cores and where the traffic characteristics of the system cannot be predicted statically [2].

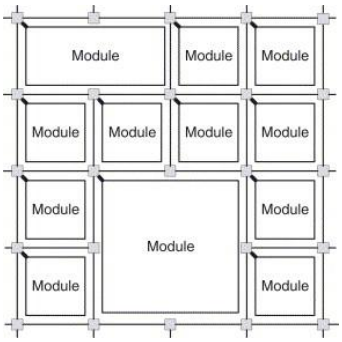

Fig. 2: NoC supporting heterogeneously sized cores

However, SoCs can also be heterogeneous, with each core having different size, functionality and communication requirements. This leads to large wiring complexity after floorplanning as well as significant power and area overhead in standard and regular NoC. In such scenario, irregular NoC custom made according the heterogeneity of the cores and communication requirement were found to be more suitable and performance concise. However components of such irregular NoC were required to be custom made according to the specific application under consideration and therefore such heterogeneous NoC were not in a position to enjoy the liberty of mass production of its components as was the case with standard 2D NoCs. The transistor packing density per chip is continually following the Moore's law. However, scaling on-chip networks having hundreds of cores in two dimensions is becoming nontrivial with the increasing number of cores per chip. The 2D NoC are becoming prone to communication bottleneck and long interconnects with

increasing cores per chip [3]. Moreover increasing chip packing density by continuously shrinking transistor feature size and reaching a range of 20 to 40 nm is throwing new challenges. As the technology reaches physical limitations the issues such as signal integrity, power dissipation, leakage power, clock distribution are progressively becoming intractable [4].

In such scenario, 3D integration technologies [5] has emerged to enable stack multiple dies on a single chip, creating 3D integrated circuits and offering an opportunity to be the next performance growth engine. 3D integrated circuits may facilitate design and fabrication of homogenous as well as heterogeneous and complex system on chips with high communication demand with efficient performance, energy efficiency, and modular design. The 3D integrated circuits are facilitating the NoC to scale over the third dimension, resulting in 3D NoCs [6]. 3D NoCs overcome the limited scalability of 2D NoCs over 2D planes by using short and fast vertical interconnects of 3D Integrated circuits. In comparison to 2D NoCs, the 3D NoCs significantly reduce the network diameter and communication distance leading to improvement of communication performance and reducing power consumption with increased network scalability.

This paper presents a brief study of NoC along with major challenges in its journey from 2D plane to 3D plane. Section 2 elaborates NoC topology parameters. Standard 2D-NoC topologies are discussed in section 3. Irregular NoC's significance and challenges are highlighted in section 4. Popular 3D NoC architectures along with design challenges are briefed in section 5 and in section 6 the conclusion is presented.

II. NOC TOPOLOGY PARAMETERS

The topology is the arrangement of nodes and channels in the NoC [7]. It establishes a communication infrastructure for the processing elements and can generally be modeled as a graph. A topology of NoC has some important characteristic parameters such as degree, diameter, link complexity and bisection width. These parameters together characterize a NoC topology and distinguish one topology from the others.

Node Degree is the number of links connected to a core/processing element. A topology is referred as regular if all its cores/nodes have the same degree. Small and fixed core degree is generally preferred as it supports modular component design and is easily scalable. Moreover there is also the physical limitation on the numbers of ports (degree) a router can support.

Diameter is the maximum shortest path between any pair of nodes in the NoC. If there is no direct connection between two nodes, the communicating information packet has to traverse vide the intermediate cores/routers which may introduce multiple hop/switching delays. In minimum routing, diameter determines the worstcase distance so the latency and so delay for the communication.

Channel Complexity is the total number of channels in the NoC. As the NoC scales, the channel complexity increases. Adding more channels to a certain network can reduce its diameter and provide better communication among the cores but increase the hardware complexity and may increase the area overhead, which is an important design parameter for the on-chip networks.

Bisection Width is *the* number of channels required to be removed to divide the topology into two equally sized networks. A large bisection width is preferable as it provides better connectivity among the sub-networks. In standard NoCs ring, star and tree are the only topologies supporting fixed bisection width.

III. 2D STANDARD NETWORKS ON CHIP

A Network on chip is characterized by its topology [7], which is basically the underlying graph, in which a switch is represented by a node and a link between two switches is represented by an edge. Various standard topologies have been proposed and used for Network on Chip. Meshes, torus, binary trees, ring and fat trees are a few examples of the topologies used for NoC. Figure 3 shows some popular NoC topologies. A standard topology is more suitable for NoC design if it has an area efficient layout. The inter-related parameters which characterize the topology and performance of a NoC are network diameter, connectivity, bandwidth and latency. Network diameter is the maximum number of intermediate nodes between a source and any destination pair. Connectivity refers to the number of direct neighbors of any switch node in the network. Bandwidth is a measure of maximum rate (in bits/second) of information flow in the network. Latency is the time taken by a message to travel from the source core to the destination core.

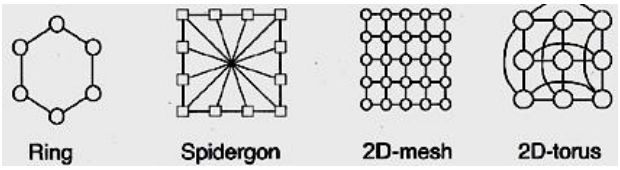

Fig. 3: Popular standard 2D NoC topologies

Some of the popular NoC topologies, such as ring, star, mesh, fat tree, and torus are briefly discussed here.

In ring topology, all nodes are connected in a ring fashion. Every node has two neighbors regardless of the size of the ring. Its small degree is preferable, but it's diameter increases linearly with the number of nodes. The star topology architecture, assume there are N nodes and the center node has N-1 neighbors. Only the center node has a degree of N-1. Other nodes only have degree of 1. Diameter of a star architecture is 2, regardless of its size. Star topology has a small diameter leading to short average hop distance but the center node can become a communication bottleneck due to congestion. In a mesh, nodes are connected as a grid. In mesh nodes have different degrees according to their locations within the mesh. The advantage of mesh topology includes easy extendibility, multiple paths for communication and fault tolerance due to link failure. However the diameter of Mesh can be large. In a fat tree, only leaves are processing elements and all internal nodes only acts as routers. The degree of nodes increases as we move upwards towards the root of the fat tree and so traffic congestion can occur towards the root in case of high communication requirements. A torus topology is obtained by adding direct links to the two end nodes in the same row or column of the mesh topology. Compared to meshes, in torus the diameter of the topology is reduced due to these wrap around links but these wrap around links can be quite long leading to communication synchronization problems. However we can have folded torus topology with reduced wire length.

Generally, mesh topology makes better use of links, while tree-based topologies are useful for exploiting locality of traffic. Routing in NoC is required to be deadlock free and topology can play a big part in deciding a deadlock free routing for NoC. Generally, for NoC turn prohibition based deadlock free routing is advocated [7]. Two such routing functions for 2D mesh based NoCs are XY [7] and OE [8]. Both these routing functions are theoretically guaranteed to be free of deadlock and livelock.

IV. 2D IRREGULAR NETWORKS ON CHIP

The standard topologies are appropriate for homogenous generic NoC architecture requiring modular design but for heterogeneous application specific NoC, standard topologies are not appropriate as in such NoC each core may be having different size, functionality and communication requirements and so standard topologies can have a layout that poorly matches application communication requirements. This can lead to large wiring complexity after floorplanning. In such scenario, an irregular topology (Fig. 2) designed according to the core sizes and/or application's communication behaviors are more apt.

In addition to above, the generic regular topologies can also become irregular for supporting clustering, oversized region or due to faults in switches or links in the regular NoCs. The region concept presented in [9] was intended for use of larger resources, which do not fit in the fixed sized slot of a regular mesh architecture layout. However, the concept can also be used for logical and physical clustering of cores and encapsulating a group of resources with specialized requirements on performance, power consumption or data security.

If the topology is regular, it is wise with regards to performance, to use a topology dependent deadlock free routing such as XY [7] and OE [8] since it would be able to exploit the regularity of the topology but such routing functions are susceptible to topology changes. A faulty switch or link can degrade the topology into an irregular topology leading to failure of such routing functions. An effective way to accomplishing fault-tolerance can be to use topology agnostic deadlock free routing functions. As deadlock-free routing is critical for proper operation of such irregular NoCs, several turn prohibition based topology agnostic deadlock free routing algorithms such as are up*/down* [10], lturn [11], down/up [12], prefix-routing [13] are proposed in the NoC research literature.

V. 3D NETWORKS ON CHIP

The 3D Integration Circuits is a cutting edge technology in which multiple 2D chip layers are stacked vertically via layer-to-layer interconnections and containing multiple layers of active processing elements. An example 3D NoC is exhibited in Fig. 4. Active Research across the world on 3D NoC is being pursued. Some of the prominent 3D NoC research [14] includes chip stacking, transistor stacking, die-on-wafer stacking, and wafer-level stacking. With the shrinking of the feature size 3D integrated circuits [15] have proved itself as an attractive option for overcoming the barriers in interconnect scaling thereby offering an opportunity to continue performance improvements using CMOS technology. 3D integration technologies offer huge potential for futuristic multicore NoC with hundreds if not thousands of processing elements on a chip. Some of the advantages of 3D NoC include reduction in interconnect wire length, reduction in the diameter of the network, increased memory bandwidth, improved form factor leading to higher packing density and smaller footprint.

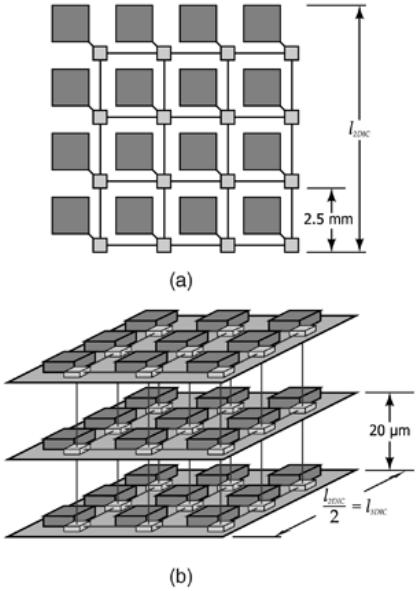

Fig. 4: 4×4 2D-NoC (a) and 4×4×4 3D-NoC (b)

Numerous 3D-NoC architectures were proposed in recent NoC research literature. Each one of these architectures have their own merits and demerits. Some of these 3D-NoC designs are briefly discussed in the following paragraphs.

A prominent 3D NoC design is *Symmetric NoC Architecture*. In this design cores are grouped into multiple layers and are stacked on top of each other. The drawbacks of such design include negligible inter-wafer distance in 3D chips [16] and larger crossbar in each switch due to increase in number of ports per switch.

3D NoC Bus Hybrid Architecture [17] was proposed as a hybrid between packet-switched network and a bus. This hybrid system provides both performance and area benefits. It requires a smaller crossbar per switch in comparision to the symmetric 3D NoC design. Hybrid architecture is also more energy and area efficient. The bus link can have its own dedicated queue, which is controlled by a central arbiter. Flits from different layers wishing to move up/down should arbitrate for access to the shared medium. However this design can not support concurrency in the 3rd dimension.

In *Ciliated 3D Mesh Architecture* [18], each switch contains at most 5+k ports (one for each cardinal direction, two for up and down (one either up or down in two layer 3D mesh) and one to each of the k IP blocks. In consequence of multiple IP cores per switch and diminished connectivity this architecture presents lower overall bandwidth compared to a symmetric 3D Mesh but is generally more energy efficient.

Tree-based 3D NoCs: Butterfly fat tree (BFT) and the generic fat tree or SPIN [19] are the two types of tree-based interconnection networks that have been considered for 2D-NoC. According to [18], considerable enhancements can be achieved when these networks are instantiated in a 3D IC environment. Unlike the work with mesh-based NoCs, any new topologies for tree-based systems were not proposed instead the [19] presents an achievable performance benefits by instantiating already existing tree-based NoC topologies in

a 3D environment. when the 2D BFT network is mapped onto a multi-layer 3D SoC, wire routing becomes simpler and the longest inter-switch wire length is reduced by at least a factor of two, in comparison with the one-layer 2D implementation. This can lead to reduced energy dissipation as well as smaller area overhead. B. S. Feero et. al [18] claim that the fat tree topology will have the same advantages when mapped onto a 3D IC as the BFT.

Similar to 2D-NoC, the turn prohibition based routing can be applied to the 3D-NoC with requisite extension and modifications. A natural extension of XY routing is the XYZ routing in 3D-NoC. In spite of all the advantages of 3D-NoC, this design paradigm presents quite a few challenges such as thermal management, fabrication issues such as coupling between TSVs, power management issues and electrical modeling challenges to name a few.

VI. CONCLUSION

The paper presents the progress in the domain of NoC design space from 2 dimensional plane to the 3 dimensional plane. The paper highlights the various challenges faced by the NoC research community and the various effective solution proposed in the domain of NoC research.

The 3D-NoC seems to be the natural extension of the 2D NoC after advent of the effective 3D integrated circuit fabrication technology. Of course with the inception of 3D NoCs, the achievable performance benefits are expected to increase drastically. However 3D NoC also enforces many new design and fabrication challenges, which require active research in this domain to be addressed.

ACKNOWLEDGMENT

This work is supported by Department of Science and Technology, jaipur, Rajasthan, India under the research project " *Network-on-Chip simulation framework for regular, irregular and 3D-Mesh Interconnection Architecture* "

REFERENCES

[1] W. J. Dally and B. Towles, "Route packets, not wires: on-chip interconnection networks," in Proceedings of the 38th conference on Design automation, June 2001, pp. 684–689.

[2] M. Taylor, W. Lee, S. Amarasinghe, A. Agarwal, "Scalar Operand Networks ", in IEEE Transactions on Parallel and Distributed Systems, vol. 16, no. 2, pp. 145-162, Feb 2005.

[3] V. Pavlidis and E. Friedman. "3-D topologies for networks-on-chip", in IEEE Transactions on Very Large Scale Integration Systems, 15(10), 2007.

[4] The International Technology Roadmap for Semiconductors (ITRS) update, Semiconductor Industry Association, 2008.

[5] G. Philip, B. Christopher, and P. Ramm, *Handbook of 3D Integration*. Wiley-VCH, 2008.

[6] L. P. Carloni, P. Pande, and Y. Xie, "Networks-on-chip in emerging interconnect paradigms: Advantages and challenges", in Proceedings of the 3rd ACM/IEEE International Symposium on Networks-on-Chip (NOCS'09), May 2009.

[7] J. Duato, S. Yalamanchili, L. Ni, Interconnection Networks : An Engineering Approach, Elsevier, 2003.

[8] G. M. Chiu, "The odd-even turn model for adaptive routing," in IEEE Transactions on Parallel and Distributed Systems, vol. 11, no. 7, pp. 729–738, Jul 2000.

[9] 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.

[10] e. a. M. D. Schroeder, "Autonet: A High-Speed Self-Configuring Local Area Network Using Point-to-Point Links", In Journal on Selected Areas in Communications, vol. 9, Oct. 1991.

[11] A. Jouraku, A. Funahashi, H. Amano, M. Koibuchi, "L-turn routing: An Adaptive Routing in Irregular Networks", In the International Conference on Parallel Processing, pp. 374-383, Sep. 2001.

[12] Y.M. Sun, C.H. Yang, Y.C Chung, T.Y. Hang, "An Efficient Deadlock-Free Tree-Based Routing Algorithm for Irregular Wormhole-Routed Networks Based on Turn Model", In International Conference on Parallel Processing, vol. 1, pp. 343-352, Aug. 2004.

[13] J. Wu, L. Sheng, "Deadlock-Free Routing in Irregular Networks Using Prefix Routing", DIMACS Tech. Rep. 99-19, Apr. 1999.

[14] R. S. Patti, "Three-dimensional integrated circuits and the future of system-on-chip designs," in Proc. of IEEE. vol. 94, pp. 1214 – 1224, June 2006.

[15] W. R. Davis, et al., "Demystifying 3D ICs: the pros and cons of going vertical," in proc. of IEEE Design and Test of Computers. Vol. 22, pp. 498– 510, November 2005.

[16] J. Kim et al., "A novel dimensionally-decomposed router for on-chip communication in 3D architectures," in Proc. of International Symposium on Computer Architecture, pp. 138-149, 2007.

[17] A.M. Rahmani, et al., "Congestion aware, fault tolerant and thermally efficient inter-layer communication scheme for Hybrid NoC-Bus 3D architectures," Networks on Chip (NoCs), Fifth IEEE/ACM International Symposium. Pittsburgh, pp. 65-72, May 2011.

[18] B. S. Feero, and P. P. Pande, "Networks-on-Chip in a three-dimensional environment: a performance evaluation," IEEE Transactions on Computers. vol. 58, pp. 32-45, January 2009.

[19] P. Guerrier and A. Greiner, "A generic architecture for on-chip packet-switched interconnections," in Proc. of Design, Automation and Test in Europe Conference. pp. 250-256, 2000.

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 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