Accurate Performance Analysis of 3D Mesh Network on Chip Architectures

Bheemappa Halavar, Basavaraj Talawar
SPARK Lab, Computer Science and Engineering, National Institute of Technology Karnataka, India.
{cs14f06.bheem,basavaraj}@nitk.edu.in

***Abstract*—With the increase in number and complexity of cores and components in CMPs and SoCs, a highly structured and efficient on-chip communication network is required to achieve high-performance and scalability. Network on Chips(NoC) emerged as the reliable communication framework in CMPs and SoCs. Many 2-D NoC architectures have been proposed for efficient on-chip communication. In this paper, we explore the design space of 3D NoCs using floorplan driven wire lengths and link delay estimation. We analyse the performance and cost of 2D and two 3D variants of the Mesh topology by injecting two synthetic traffic pattern for varying buffer space and floorplan based delays were considered to for the experiments. Results of our experiments show that for the injection rates from 0.02 to 0.2 the average network latency of a 4-layer 3D Mesh is reduced up to 54% compared to its 2D counterpart. The on chip communication performance improved up to 2.2× and 3.1× in 4-layer 3D Mesh compare to 2D Mesh with uniform and transpose traffic patterns respectively.**

***Index Terms*—3-D integration, Network-on-chip (NoC), Through-silicon via (TSV), Interconnect, 3D topologies, Design space exploration**

I. Introduction

Use of Network on Chips(NoCs) in Chip-Multiprocessors (CMPs) and System-on-chips(SoCs) as the communication fabric leads to better scalability and performance than the conventional on-chip communication fabric [1]. NoC is the on-chip interconnect paradigm in which chip modules (called nodes) exchange data through an optimally designed packet routing network instead of conventional wires. The packets are split into flow control units called flits and are routed between source and destination nodes through a series of routers and links [2]. The routers of the NoC are connected in a well defined topology driven by the bandwidth requirements of the application on CMPs and SoCs [3].

3D NoCs have lesser aggregate wire length resulting in improved communication latency and power compared to their 2D counterparts [4]. It has been estimated that 3D architectures reduce wiring length by a factor of the square root of the number of layers used [5].

Early stage estimation of power and performance metrics like latency, power and energy through cycle accurate simulators is critical in reducing the time spent in the NoC design cycle. These simulators explored various architectural and micro architectural design parameters such as router pipeline depth, arbitration techniques, number and size of virtual channels, number and size input/output buffers, routing and switching techniques, link latency, link width, network partitions, node concentrations, express physical links[6].

Current simulators consist of fixed delay component and the link length is driven by the physical dimension of the components on the chip. Incorporating physical characteristics of the chip is necessary to estimate the correct length of the link. Hence depending on the topology, links of varying length and latency exist in the chip. Accurate estimation of individual link delay requires considering the length, the operating frequency and delay of the interconnect. Selection of topology is a critical NoC design decision. The topology decision is based on the bandwidth requirements of the application, resource available, area budget, power and performance constraints of the chip. The most common topology used is the Mesh because of its symmetric nature and short wire lengths between routers[7].

In this work, the design space parameters explored are, number of VCs(V) and the depth of the buffers(D). The overall performance of NoC depends on the delay in the links and routers. To get accurate performance, Floorplan is carried out by considering physical characteristics of the PEs and routers. Based on the floorplan, accurate link delay is estimated using the link and TSV delay model. Further, these delays are used in simulation to obtain the correct performance of the NoC architecture. The main focus of this paper is the detailed comparative evaluation of 3D-NoC architectures with their 2D counterparts to determine the performance benefits quantitatively. Performance and cost trade-offs is evaluated for conventional 2D Mesh with 2-layer 3D Mesh, 4-layer 3D Mesh NoC topologies considering different V and D and NoC topology has been characterized in the presence of uniform and transpose traffic patterns using cycle-accurate simulation.

A. Contribution of This Work

1) Extension and analysis of existing 2D Mesh topology to 2-layer and 4-layer in 3D by considering TSV as a vertical connection in the third dimension. Accurate link length derived from the respective floorplans.

978-1-5386-1112-8/18/$31.00 ©2018 IEEE

Delay estimation by employing ORION delay models.

2) Analysis of buffer space utilization among the topologies 2D and 3D topologies.

The paper is organized as follows. Related work has been presented in Section II. Floorplan of 2D and 3D topologies with accurate physical characteristics values has been detailed in Section III. Section IV discusses the link delay and buffer space analysis. Experimental set up has been described in Section V. Results have been presented in Section VI. The paper concludes in Section VII.

II. Related Work

A. Design Space Exploration

Pande et al. [8] discussed the evaluation methodology to compare the performance and characteristics of a SPIN, Torus, Folded torus, Octagon, BFT NoC architectures. Higher throughput and lower latency were observed for higher degree of connectivity. At saturation, SPIN and Octagon have higher average energy dissipation than the other NoC architectures. The switches were designed by considering four virtual channels to maintain low latency and considerable throughput. In this paper, only 2D architectures have been evaluated.

Designers have explored the architectural design space to improve performance and energy-efficiency. Architectural choices for NoCs consists of express physical links, network partitioning and concentration. Alternative NoC architectures have been proposed based on design options like PEs concentration, Multiple physical networks, virtual channels, express channels, flit widths. These design options are evaluated for the energy throughput ratio (ETR) metric [6][9].

Physical and virtual express topologies are proposed to address the scalability issue in 2D Mesh. Chen et al. [10] have compared the Mesh based physical and virtual topologies with an increased number of nodes. Physical express topologies give better throughputs and also it's more robust compared to virtual express topology.

Feero et al. [11] discuss 3D topologies derived from Mesh and Tree topologies. Latency, energy dissipation, and wire area overhead of 3D NoC architectures are compared with 2D NoC architectures. With small area overhead Mesh-based architectures have significant performance gain. Latency characteristics better for 3D Mesh over the 2D Mesh. 3D tree-based NoCs have area overhead and significant gain in energy dissipation. In NoC, accurate link latency also play important role in overall performance. In this paper accurate link latency and vertical link delay has not been considered for evaluation.

With use of 3D NoC partitioning, 3D Stacked Mesh and 3D Stacked Hexagonal topologies are generated and analyzed by comparing their performances with Stacked 2D-Mesh NoC and classical 2D- Mesh and 3D-Mesh NoC. 3D NoC architectures using partitioning method show better performance than regular 3D Mesh topology[12].

On-chip simulators consider a static wire length or constant delay on the communications. But the length and the delay of NoCs link vary according to the floorplan of the NoC. In this work, 2D and 3D variants of Mesh topologies are considered for experimental analysis based on floorplans of each the topologies. The floorplan is used to calculate the exact length and number of horizontal and vertical links. The link delay models from ORION are used in the experiments.

For evaluation of NoC architecture, BookSim cycle-accurate simulator has been considered. Power of the NoC components and accurate link delay with the optimal (number, size) repeater is calculated for the individual length of the wire using ORION models. The cycle accurate link delays were updated in the Booksim. Further, network models were modified in Booksim to model the behavior of the 3D variants.

III. Floorplan of 2D & 3D NoC Topology

A. 2D Mesh Topology

Mesh is direct network topology which allows integration of more PEs in a regular shape structure [13], each routers are connected to all its neighbouring routers. The floorplan of 64 node 2D Mesh used in this work is shown in Figure 1.

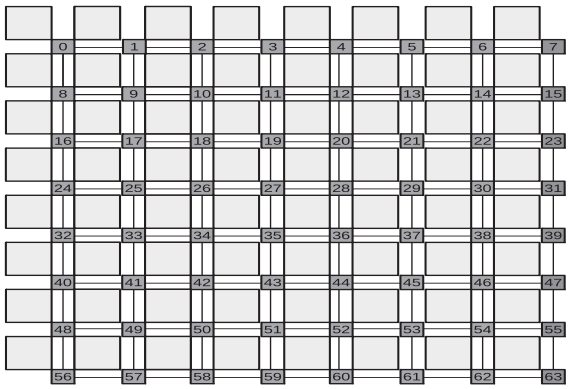


Fig. 1: Floorplan of 2D Mesh with 64 PEs. Wire length is constant through out the floorplan and 128bit channel.

The floorplan consists of a system with tiled Chip Multiprocessor with 64 Sun-SPARC cores [14] and area of core 3.4mm^2. Router area is estimated from ORION3.0 and which shows 5 port router area is 0.598509mm^2 and 7-port router area is 0.865314mm^2. Table I depicts the micro-architectural parameters used.

B. 3D Mesh Topology

Existing Mesh topology in cycle accurate simulator works for k array, n cube. Where is k is radix(number of elements in each dimension) and n is the number of dimensions. For example, 8×8 Mesh topology is `8-array, 2-cube` 2D Mesh topology has total number of routers

TABLE I: Parameters used in the design of the floorplan

Parameter	Value
Technology	32nm
Clock Frequency	2.5GHz
PEs area	3.4mm^2
4-port router area	0.47098mm^2
5-port router area	0.598509mm^2
7-port router area	0.865314 mm^2
Channel size	128 bit
TSV Delay	1 Clock cycle

(k^n)=64. 3D Meshes aim to reduce this latency by redistributing nodes vertically and hence the 64 8×8 Mesh can be converted into two 3D configuration as $4 \times 4 \times 4$ and $8 \times 4 \times 2$. Two 3D NoC topologies are designed from existing 2D Mesh topology to analyse the performance of going vertical. Figure 2 and 3 show the floorplan based architecture of $8 \times 4 \times 2$ and $4 \times 4 \times 4$ 3D Mesh topology and both the topology have 32 and 16 routers per layer respectively. TSVs are used to connect interlayer routers and delay of the TSVs is considered as 1 clock cycle [15]. The horizontal link delay models are derived from ORION and are used to calculate accurate horizontal wire delay.

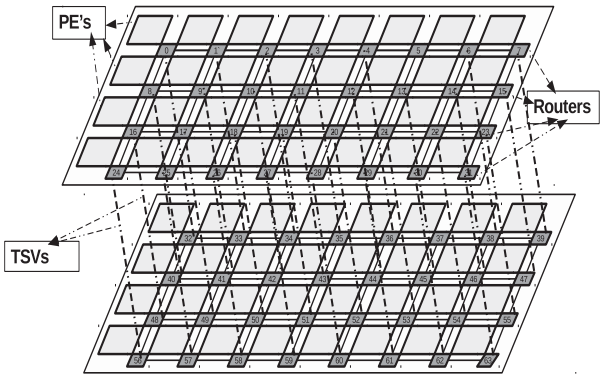


Fig. 2: 8×4×2 3D Mesh with four stacked layers connected using TSVs. Each are 128 bit TSVs.

IV. Link delay and Buffer Space Analysis

A. Horizontal link and Vertical link Delay Estimation

This section presents the delay estimation of horizontal and vertical links. Link lengths are extracted from the floorplan of the topologies. Table II shows the delays details about the horizontal and vertical link analysis of 3 topologies. The second column in Table II shows the wire length and the delay of wire for 2.5GHz frequency is calculated using RC delay models of ORION2.0. Delays and link counts of both HL and VL(TSV) are shown

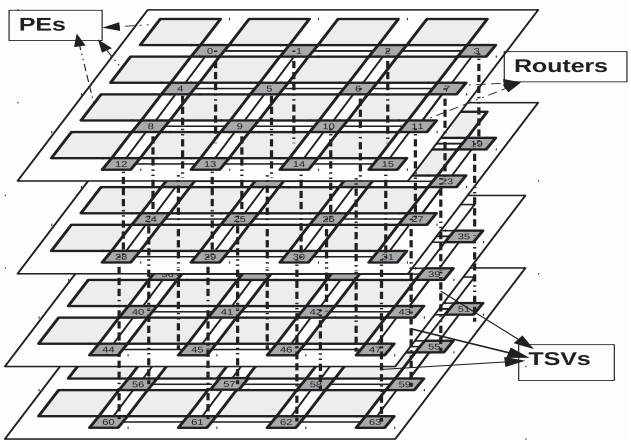


Fig. 3: 4×4×4 3D Mesh with four stacked layers connected using TSVs. Each are 128 bit TSVs.

in Table II. For the experiments the TSV details are considered from [16], the diameter is 5μm with 10μm pitch while TSV length is 20μm after wafer thinning. The vertical connections are TSVs and 128 bit parallel TSV connection because in simulation HL channel size is 128 bit is considered for higher bandwidth.

TABLE II: Horizontal link(HL) length and delay details of 2D and 3D variants of Mesh. These delays are considered for the simulation

Topology	Wire(mm)	HL (wire) count	HL Delay (clock cycle)	VL count	TSV count	Delay of TVS (clock cycle)
2D Mesh	1.844	112	4	-	-	-
2-layer 3D Mesh	1.844	108	4	32	4096	1
4-layer 3D Mesh	1.844	96	4	48	6144	1

B. Buffer Space Analysis

This section discusses the buffer space used for the 2D and 3D variants. The overall buffer space of the topology rely on the total number of routers, input/output ports and virtual channels per port, buffer depth per VC of a topology. Based on these parameters total buffer space utilization(B) is represented in Equation 1.

$$B = \sum_{i=1}^{n} (R_i * P_i) * V * D \quad (1)$$

Where n is the types of routers of each number of i/o ports, R_i is the total number of routers in i^{th} category, total number of ports in R_i is P_i, number of Virtual channels per port is V, D is Buffer depth per VC. B is the total buffer space in terms of the number of flits for the topology.

Table III shows the total buffer space utilized for the 2D and 3D Mesh topology for different VC(V) and VC

TABLE III: Buffer Space of 2D and 3D topologies

VC parameters		Buffer Space (flits)		
V	D	Mesh	2-layer 3D-Mesh	4-layer 3D-Mesh
4	4	5120	6144	6656
	8	10240	12288	13312
6	4	7680	9216	9984
	8	15360	18432	19968
(R_n,P_n)		(64,5)	(64,6)	(32,6) (32,7)

depth(D). In Table III, first two columns are the virtual channel(VC) parameters (number of VC and VC depth) and remaining column consist of total buffer space for the 2D and 3D Mesh topology. The last row of Table III represents the number of routers and ports (R_n, P_n) present in the routers. 4-layer 3D Mesh consists two different types of routers with 6-port and 7-port (n=2). For example, if VC=6, D=8, 4-layer 3D Mesh has 32 6-port routers and 32 7-port routers, so the total buffer space is 19968 as shown in the Table III. Each topology has different buffer space based on the V and D. 4-layer 3D Mesh has higher buffer space compare to 2D Mesh topology.

V. Experimental Setup

Table IV shows network configuration parameters of 2D and 3D Mesh topologies. Cycle accurate simulator is modified to evaluate 2D and 3D variants of Mesh topologies with uniform random and transpose traffic pattern. We have considered 128 flits per channel and packet length is of 5 flits. Variable VCs and VC depths are considered to check the which topology supports higher injection rate, better bandwidth and performance.

TABLE IV: Simulated Network Configuration.

BookSim Parameter	Value
Topology	2D Mesh & 2-layer 3D Mesh & 4-layer 3D Mesh
Network Size	64 Nodes
Switches	64
Traffic	Uniform Random, Transpose
Number of VCs	4,6,8
VC buffer size	4,8,12,16
Simulation time	10^5 cycles

The cycle-accurate on-chip network simulator(BookSim2.0) is modified to support 2-layer 3D and 4-layer 3D NoC with accurate delay and used in our experiments. For the 2D and 3D Mesh, XY and XYZ routing is used. Network modules in Booksim were modified to give the correct routing functionality for the 2-layer and 4-layer 3D Mesh experiments. Link lengths are estimated based on the floorplan as shown in Figure 1, 3 to get the accurate performance metric. Delays of the horizontal links in each topology are estimated and plugged in the simulator as detailed in Section III.A and the TSVs delays were modeled from existing works [15].

VI. Results and Discussion

A. Average Network Latency

1) 2D and 3D Mesh Topology: Use of floorplan based latencies for links are essential for accurate calculation of the communication time. The deviation from the actual average latency in Booksim is shown in Figure 4.

Over various injection rates, the average network latency obtained from Booksim using default link latencies and from floorplan based link latencies are plotted in Figure 4. When we use the floorplan values in simulation, an increase in average network latency from 19% to 43% observed. Floorplan and accurate delay estimation is essential before the simulation to obtain the correct performance of the NoC architecture. In a Mesh, variation in latencies is small because of the uniform link length, but as the length of the link varies, average network latency is affected significantly.

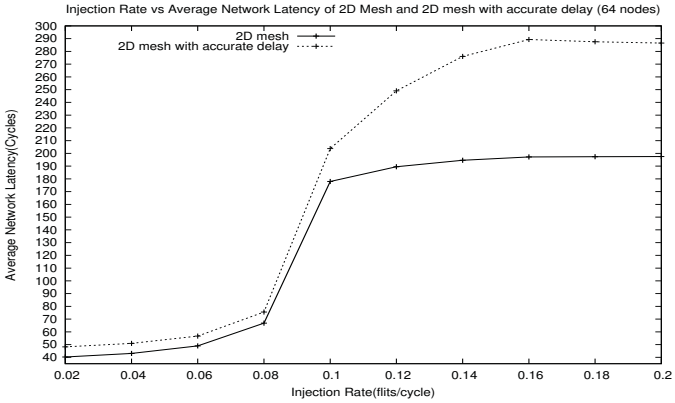
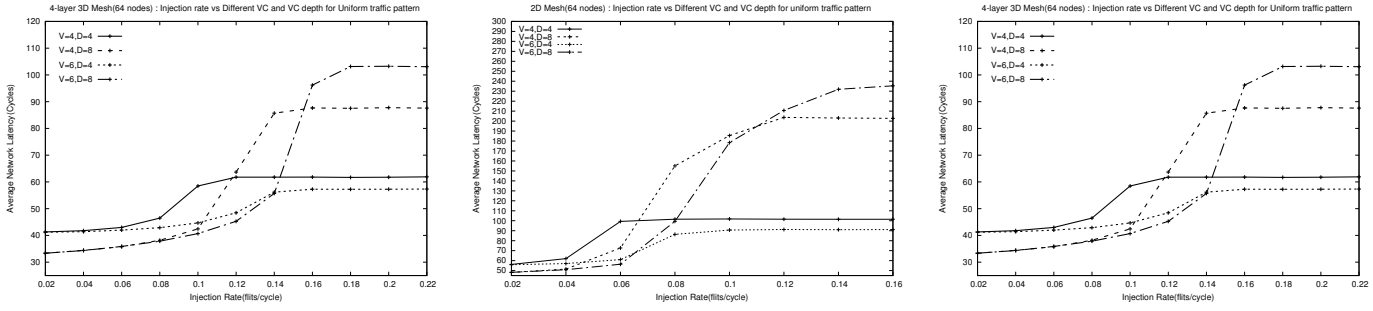


Fig. 4: Average network latency comparison for 2D Mesh and 2D Mesh with accurate delay.

2D and 3D Mesh variants topologies are observed for two different buffer variations. one is for varying VC depth (D) and other is for varying VCs (V). Figure 5 shows the latency graph of Uniform traffic pattern with different V and D. Figure 6 shows the latency graph of transpose traffic pattern with different V and D.

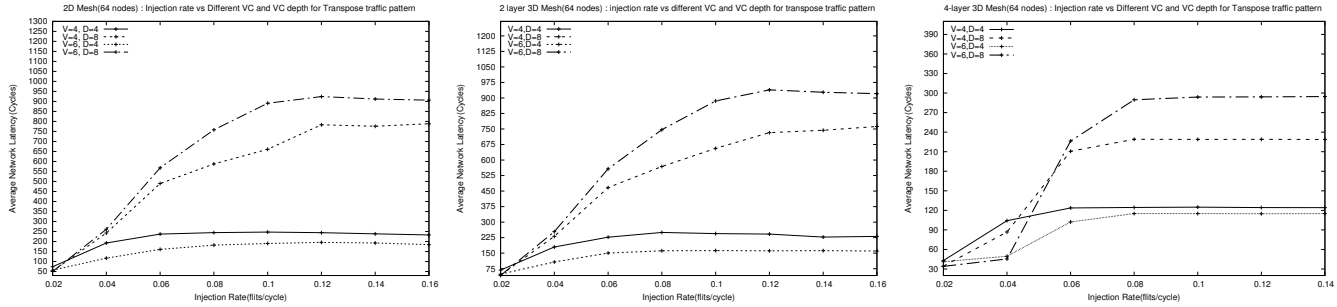
Table V depicts the simulation results which are based on the Figure 5 and 6. Table V shows the saturation points for each traffic pattern with the increase in the number of VCs and VC depth. In Table V, the saturation point is constant after 8 buffer depth (D) in both traffic pattern. Buffer Depth depends on packet size and in the simulation we have considered five flits per packet. Thus after D=8, the saturation is constant.

For Varying VC for D=8, the network saturation point increases as the number of V increases and improved



(a) (b) (c)

Fig. 5: Average network latency comparison for different VC and VC depth of for transpose Uniform pattern (a) 2D Mesh topology (b) 2-layer 3D Mesh topology (c) 4-layer 3D Mesh topology



(a) (b) (c)

Fig. 6: Average network latency comparison for different VC and VC depth of for transpose traffic pattern (a) 2D Mesh topology (b) 2-layer 3D Mesh topology (c) 4-layer 3D Mesh topology

TABLE V: 2D and 3D Mesh variants network saturation details for different buffer space based on Figure 7 and 6

Topology	Varying D (Buffer Depth)				Varying V (VC)			
	V	D	S.P (Uniform)	S.P (Transpose)	D	V	S.P (Uniform)	S.P (Transpose)
2D Mesh	6	4	0.12	0.8	8	4	0.14	0.12
		8	0.16	0.14		6	0.16	0.14
2-layer 3D Mesh	6	4	0.12	0.12	8	4	0.14	0.14
		8	0.16	0.14		6	0.16	0.16
4-layer 3D Mesh	6	4	0.16	0.1	8	4	0.16	0.12
		8	0.2	0.14		6	0.2	0.14

network throughput and thus, NoC supports higher injection rate with increase in V. From figure Comparison of latencies between the Uniform and Transpose traffic pattern shows a difference of up to 230% for 4-layer 3D mesh. The uniform distribution of the traffic results lower in contention in the links compared to the transpose traffic pattern.

2) 2D Mesh vs 3D Mesh: Figure 7 depicts the comparison of latencies for 2D Mesh, 2-layer 3D Mesh and 4-layer 3D Mesh for uniform and transpose traffic pattern with V=8 and D=8. 4-layer 3D Mesh with uniform traffic has a reduction in the average network latency up to 54% compare to 2D Mesh uniform random traffic throughout the simulation. We observed that the decrease in the average network latency from upto 68% in 4-layer 3D Mesh compared 2D Mesh with transpose traffic pattern. The latency is decrease because there is a reduction in horizontal links from 112 to 96 in 4-layer 3D Mesh and there 48 extra VL as shown in resources usage Table VI.

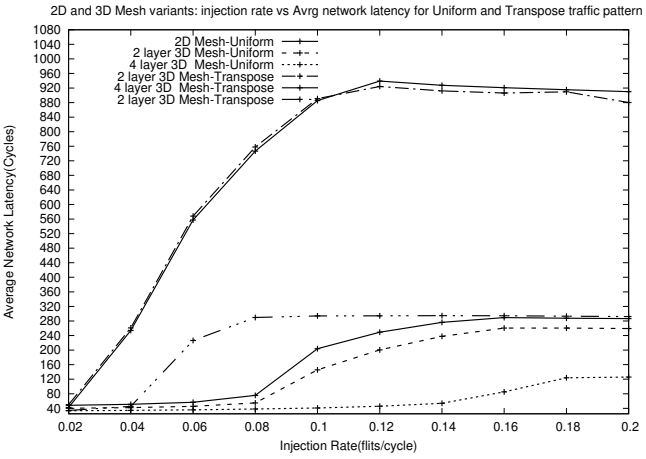


Fig. 7: Average Network latency comparison between all six topology i.e both 2D and 3D version of Mesh, BFT. Mesh 4-layer shows better performance than among all other topology.

Experiments results show that the 4-layer 3D Mesh has up to 2.2× on-chip commination performance improve-

ment compared to 2D Mesh with uniform traffic pattern. For transpose traffic pattern, the 4-layer 3D Mesh has up to 3.1× performance improvement compared to 2D Mesh. There are two reason to decrease in latency in 4-layer 3D Mesh, one is vertical links, and another one is total buffer space. TSV delays are up to 75% less compare to wire delay as shown Table II and From Table III total buffer space in the 4-layer 3D mesh is 22% more than the 2D and 8% more than the 2-layer 3D mesh.

Table VI shows the total number of resources used in each topology. From Table VI, the 4-layer 3D mesh has 6 and 7 port routers, 32 extra links compare to 2D mesh and 16 additional vertical links compared to 2-layer 3D mesh.

TABLE VI: Total number of resources used for all six topology and the network size of 64 PEs and the Links are classified as horizontal links(HL) and vertical links (VL).

NoC Topology	Network (x/k,y/n,z)			Router (In,Out,VC)				Link counts HL,VL	
	X	Y	Z	No. Router	In/Out	VC	D	HL	VL
2-D Mesh	8	8	1	64	5/5	6	8	112	0
2-layer 3D Mesh	8	4	2	64	6/6	6	8	108	32
4-layer 3D Mesh	4	4	4	64	6/6, 7/7	6	8	96	48

VII. CONCLUSION

In this paper, to get more accurate latency values, we have considered micro-architectural characteristics of on-chip networks such as the floorplan based wire lengths, and link latencies. We compared the 2D Mesh, 2-layer 3D Mesh, and 4-layer 3D Mesh by considering uniform and transpose traffic patterns through cycle-accurate simulation. Accurate wire delays were obtained using ORION delay models. Results of our experiments show that the average network latency of a 4-layer 3D Mesh is 25% to 54% lesser than its 2D counterpart for injection rates of up to 0.2 for uniform traffic pattern. The 4-layer 3D Mesh shows up to 2.2× improved on-chip communication performance compared to 2D Mesh for uniform traffic pattern. For transpose traffic pattern, the 4-layer 3D Mesh shows up to 3.1× improved on-chip communication performance compared to 2D Mesh. The 3D Mesh shows better performance over the 2D Mesh whereas it should be noted that links, router input ports, buffers are greater in the 3D Mesh compared to the 2D Mesh. In 2D and 3D Meshes, uniform traffic gives better performance compared to transpose traffic pattern. Our work suggests that accurate performance is observed only with the use of floorplan based approach and 4-layer 3D performs better compared to 2 layer 3D NoC and 2D mesh. This study can be extended by considering more microarchitecture parameter, TSV thermal issue and power metrics for further comparison between these topologies.

References

[1] W. J. Dally and B. Towles, "Route packets, not wires: On-chip interconnection networks," in *Design Automation Conference, 2001. Proceedings.* IEEE, 2001, pp. 684–689.

[2] W. Dally and B. Towles, *Principles and Practices of Interconnection Networks.* San Francisco: Morgan kaufmann, 2004.

[3] B. Grot, J. Hestness, S. W. Keckler, and O. Mutlu, "Express cube topologies for on-chip interconnects," in *High Performance Computer Architecture, 2009. HPCA 2009. IEEE 15th Int. Symp. on.* IEEE, 2009, pp. 163–174.

[4] V. F. Pavlidis and E. G. Friedman, "3d topologies for networks-on-chip," *IEEE Trans. Very Large Scale Integr. Syst.*, vol. 15, no. 10, pp. 1081–1090, Oct. 2007.

[5] J. W. Joyner, P. Zarkesh-Ha, and J. D. Meindl, "A stochastic global net-length distribution for a three-dimensional system-on-a-chip (3d-soc)," in *ASIC/SOC Conference, 2001. Proceedings. 14th Annual IEEE International.* IEEE, 2001, pp. 147–151.

[6] A. Psathakis, V. Papaefstathiou, N. Chrysos, F. Chaix, E. Vasilakis, D. Pnevmatikatos, and M. Katevenis, "A systematic evaluation of emerging mesh-like CMP NoCs," in *Architectures for Networking and Communications Systems (ANCS), 2015 ACM/IEEE Symp. on.* IEEE, 2015, pp. 159–170.

[7] 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," in *Proc. IEEE Computer Society Annual Symp. on VLSI. ISVLSI 2002*, 2002, pp. 105–112.

[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, no. 8, pp. 1025–1040, 2005.

[9] A. Psathakis, V. Papaefstathiou, M. Katevenis, and D. Pnevmatikatos, "Design space exploration for fair resource-allocated noc architectures," in *Embedded Computer Systems: Architectures, Modeling, and Simulation (SAMOS XIV), 2014 International Conference on.* IEEE, 2014, pp. 141–148.

[10] C. H. O. Chen, N. Agarwal, T. Krishna, K.-H. Koo, L.-S. Peh, and K. C. Saraswat, "Physical vs. virtual express topologies with low-swing links for future many-core nocs," in *Proceedings of the 2010 Fourth ACM/IEEE Int. Symp. on Networks-on-Chip*, ser. NOCS '10, 2010, pp. 173–180.

[11] B. S. Feero and P. P. Pande, "Networks-on-chip in a three-dimensional environment: A performance evaluation," *Computers, IEEE Transactions on*, vol. 58, no. 1, pp. 32–45, 2009.

[12] M. H. Jabbar, D. Houzet, and O. Hammami, "Impact of 3d ic on noc topologies: A wire delay consideration," in *2013 Euromicro Conference on Digital System Design*, 2013, pp. 68–72.

[13] S. Pasricha and N. Dutt, "Chapter 12 - Networks-On-Chip," in *On-Chip Communication Architectures*, ser. Systems on Silicon, S. Pasricha and N. Dutt, Eds. Burlington: Morgan Kaufmann, pp. 439–471.

[14] T. C. Xu, P. Liljeberg, and H. Tenhunen, *A Greedy Heuristic Approximation Scheduling Algorithm for 3D Multicore Processors.* Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 281–291.

[15] P. M. Yaghini, A. Eghbal, S. S. Yazdi, N. Bagherzadeh, and M. M. Green, "Capacitive and inductive tsv-to-tsv resilient approaches for 3d ics," *IEEE Transactions on Computers*, vol. 65, no. 3, pp. 693–705, March 2016.

[16] G. Katti, M. Stucchi, K. D. Meyer, and W. Dehaene, "Electrical modeling and characterization of through silicon via for three-dimensional ics," *IEEE Transactions on Electron Devices*, vol. 57, no. 1, pp. 256–262, 2010.