Floorplan Based Performance Evaluation of 3D Variants of Mesh and BFT Networks-on-Chip

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

***Abstract*—Network on Chips(NoC) emerged as the reliable communication framework in CMPs and SoCs which enables in increase the number and complexity of cores. Many 2-D NoC architectures have been proposed for efficient on-chip communication. Cycle accurate simulators model the functionality and behavior of NoCs by considering micro-architectural parameters of the underlined components to estimate performance metric. Using 3D IC technology in NoC can lead to improved communication latency and power compared to their 2D counterpart with use of through-silicon via (TSVs) as vertical interconnect. In this paper, we explore the design space of 3D variants of the Mesh and Butterfly Fat Tree(BFT) NoCs using floorplan driven wire and TSV lengths. Analysed the performance of 2D and 3D variants of the Mesh and BFT topologies by injecting uniform traffic pattern. Results of our experiments show that, average network latency of a 4-layer 3D Mesh shows better on-chip communication performance compare to other 3D variants. In 4-layer 3D Mesh, on-chip communication performance is improved up to 2.2× compare to 2D Mesh and 4.5× compare to 4-layer 3D BFT.**

I. Introduction

Three-dimensional integrated circuits (3D ICs) is emerging as a promising technology for Systems-on-Chip (SoCs). As compared to 2D designs, 3D ICs have improved performance. 3D ICs distribute logic and memory in stacked layers and NoCs interconnect these layers using direct vertical interconnects called Through Silicon Vias (TSVs)[1]. 3D NoCs have lesser aggregate wire length resulting in improved communication latency and power compared to their 2D counterparts [2]. It has been estimated that 3D architectures reduce wiring length by a factor of the square root of the number of layers used [3].

The time spent during NoC design cycle is reduced by using cycle accurate simulators as early stage estimation. Using cycle accurate simulator power and performance metrics like latency, power and energy can be estimated. Existing simulators explored various architectural and micro architectural design parameters at routers level and network level. Using these simulator we can explore network partitions, node concentrations, express physical links [4] of NoC.

The topology selection decision is based on the bandwidth requirements of the application, resource available, area budget, power and performance constraints of the chip. Mesh topology symmetric nature and has short wire lengths between routers hence its most widely used topology[5]. Butterfly Fat Tree (BFT) is another topology which has less routers and links compare to Mesh topology. Link latency plays a significant role in overall performance of NoC. Our experiments on 64 node 2D Mesh with uniform random traffic pattern shows that on an average 15-20% of flit traversal time is spent in the links of NoC for an injection rate of 0.1. However, wire lengths estimated using floorplans of the topology, we can see that the wires have varying lengths. These wires have delays greater than one cycle (shown in the paper later). Accurate modeling of link delay is necessary for the early stage of the design trade-off studies.

In this work, we evaluate and analyse the candidate topologies for performance and cost trade-offs by extending the 2D simulator support for 3D variants of the Mesh and BFT topologies based on floorplan. The main focus of this paper is the detailed comparative evaluation of 3D-NoC architectures with their 2D counterparts to determine the performance benefits and resource used. We evaluated the conventional 2D Mesh with 2-layer 3D Mesh, 4-layer 3D Mesh, 2D BFT, 2-layer 3D BFT and 4-layer 3D BFT. Performance of both 2D and 3D variants of Mesh and BFT NoC topologies have been characterized in the presence of uniform traffic patterns through cycle accurate simulation by considering accurate vertical and horizontal link delay which are derived from the floorplans.

A. Contributions of this work

1) Extended simulator to support 3D variants of the Mesh and BFT topologies with TSV as vertical connection.
2) Accurate performance evaluation and analysis of 2D Mesh, BFT topology to 2-layer and 4-layer in 3D Mesh, BFT designs based on floorplans with TSV as vertical connection in the third dimension.

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

978-1-5386-3821-7/18/$31.00 © 2018 IEEE

II. Related Work

In [6], a new 3D topological NoC design discussed based on the BFT topology with an efficient table based uniform routing algorithm for 3D NoC and the new BFT is compared with Mesh. The accurate wire delays are not considered in the simulation .

Feero et al. [7] 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 [8], proposed 3D-HiCIT (Hierarchical Crossbar-based Interconnection Topology) network-on-chip (NoC) and compared with other hierarchical topologies in terms of flexibility, scalability and performance. 3D-HiCIT allows reducing the average latency up to 50% when compared to the 3D-BFT and 45% when compared to 3D-SPIN.

In [9], for 3D NoC Bus Hybrid architecture authors have proposed an ultra-optimized inter-layer communication. There is a significant area, power, and performance improvements under uniform, hotspot of 10% compared to 3D NoC-Bus Hybrid Mesh architecture. In [10], to avoid very long global wires for Clos NoC (CNOC) 3D integration employed by 3D partitioning and floorplanning. 3D CNOC topology scales very well with an upper bound on power consumption.

On-chip simulators considers constant delay on the communications but length and the delay of NoCs link vary according to the floorplan of the NoC. In this work, 2D and 3D variants of Mesh, BFT topologies are considered for experimental analysis. Two 3D variants per topology are generated and which is of 2 and 4 stacked layers to check the behaviour of topology in 3D. Floorplans of Mesh and BFT topologies are used to calculate the exact length and number of horizontal and vertical links. The link delay models from ORION[11] are used in the experiments. For evaluation of 2D and 3D variants of NoC architectures, BookSim cycle accurate[12] simulator have been considered and modified to support the behaviour of the 3D variants with accurate delay.

III. Floorplan and Delay estimation of Mesh and BFT Topology

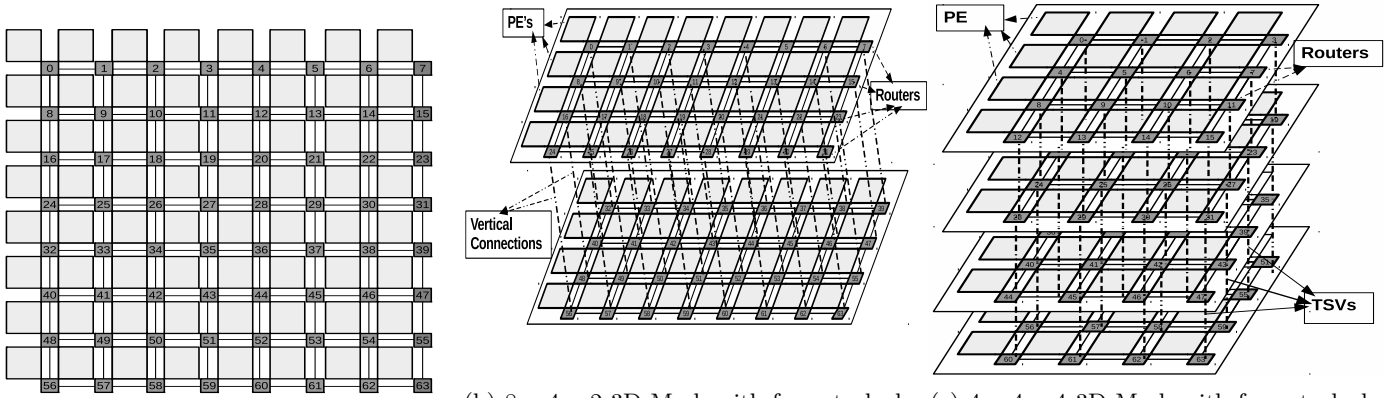
A. 2D and 3D Mesh Topology

Mesh is direct network topology which allows integration of more PEs in a regular shape structure [13], where every router except those at the edges is connected to all its neighbouring routers. 8×8 Mesh topology is `8-array, 2-cube` 2D Mesh topology has total number of routers $(k^n)=64$ as shown in Figure 1 (a). 3D Meshes aim to reduce this latency by redistributing nodes vertically. The 8×8 Mesh is converted into 2-layer and 4-layer 3D Mesh and are $4 \times 4 \times 4$ 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 1 (b) and 1 (c) show the floorplan based architecture of $8\times4\times2$ and $4\times4\times4$ 3D Mesh topology. Both the topology have 32 and 16 routers per layer respectively. TSVs are used to connect interlayer routers. The delay of the TSVs is considered as 1 clock cycle [14].

TABLE I: Parameters used in the design of the floorplan

Parameter	Value
PEs area	3.4mm^2
4-port router area	0.69mm^2
5-port router area	1.25mm^2
6-port router area	1.57mm^2
7-port router area	1.91 mm^2
Channel size	128 bit

The floorplan consists of a system with tiled Chip Multiprocessor with 64 Sun-SPARC cores [15] and area



(a) 8×8 2D mesh with 4 PEs

(b) $8 \times 4 \times 2$ 3D Mesh with four stacked layers connected using TSVs.

(c) $4 \times 4 \times 4$ 3D Mesh with four stacked layers connected using TSVs.

Fig. 1: Floorplan of 2D and 3D mesh topology. Links and TSVs are 128bit wire.

of core 3.4mm^2. Router area is estimated from ORION. All the micro-architectural parameters used are shown in Table I.

B. BFT topology

In BFT topology, PEs are placed at the leaves and routers placed at the top and intermediate levels. A pair of coordinates is used to label each node, (L, P) where L denotes a nodes level and P denotes its position within that level. BFT has different non-uniform links in each level [16].

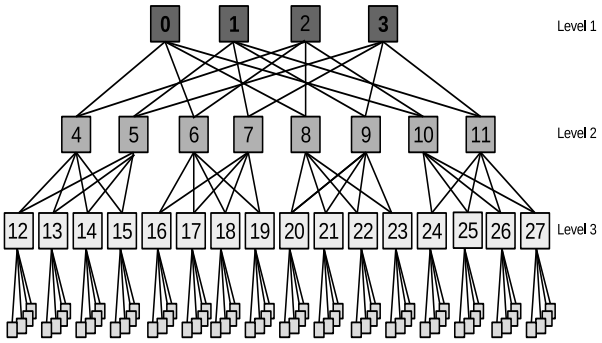


Fig. 3: 64 node BFT topology with three levels. Level 1 is of 4 router, level 2 of 8 routers, level 3 of 16 routers. The leaves are the PEs which are connected to Level 3 routers and 4 PE's per router.

1) 2D BFT: 2D BFT topology shown in Figure 3 consist of 64 PEs and 28 routers. Except for the top level routers(which have 4 ports), all routers contains 6 ports. In six port router, one port is connected to each of four child nodes and remaining two ports connected to parent nodes. Figure 2 (a) shows floorplan of 2D BFT topology. Micro-architectural parameters used for our experiments is shown in Table I. From the floorplan, there are five different links lengths and Table II depicts the lengths and their respective delay.

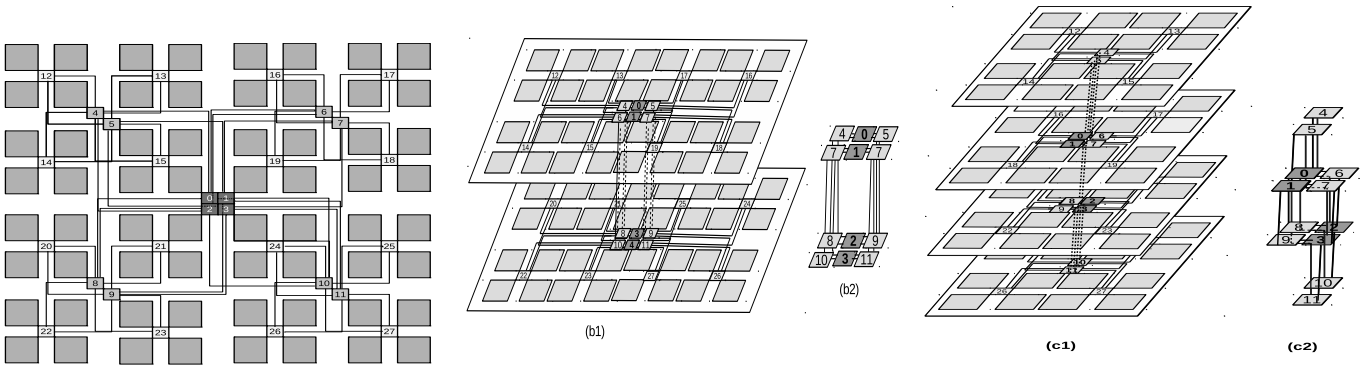
2) 2-layer 3D BFT: The level 2 routers are move towards level 1 to reduce the link length and also to connect links from level 1 routers to level 2 between the layers. Figure 2 (b) shows the 2-layer 3D BFT which is extended from 2D BFT. 2-layer 3D BFT has two stacked layer connected through vertical TSVs. In 3D topology length of wire is reduced compare to 2D BFT because of vertical connection and are shown in Table II. Figure 2 (b1) shows the vertical connections of 4-layer 3D BFT and there are 8 links which are connected vertically.

3) 4-layer 3D BFT: Figure 2(c) shows 2-layer 3D BFT modified to 4-layer 3D BFT topology with each layer consist of 16 PE's each. Level 1 routers are placed in between layers to reduce the length of TSV. Figure 2(c1) shows the vertical connection between the routers from level 1 to level 2.

C. Link Delay Estimation

Table II shows the delays details about the horizontal and vertical link analysis of both 2D and 3D variants of Mesh, BFT topologies based on flooplan. Second column in Table II shows the wire length and the delay of wire is calculated using RC delay models of ORION for 2.5GHz frequency. Both 2D and 3D variants of Mesh, BFT topologies delay and length of wire are shown in Table II. Last two columns of the Table II shows the TSVs count and TSVs delay.

The link lengths are not changed in Mesh when we move from 2D to 3D and In 3D Mesh, for each router there are two extra vertical connection are added. In BFT, some of link lengths are reduced while moving from 2D to 3D and its because of some of horizontal links are converted vertical connection(TSVs). In 2-layer 3D BFT, total 8 links count is reduced to half compare to 2D BFT and are shown in Table II and similarly in 4-layer 3D BFT. The TSV details are considered from [17], the diameter is 5μm with 10μm pitch while TSV length is 20μm for delay calculation.



(a) Floorplan of 2D Mesh with 64 PEs and each PEs are connected routers for inter PEs comminication.

(b) (b1) 8 × 4 × 2 3D Mesh with four stacked layers connected using TSVs. (b2) Inter-layer connections

(c) (c1) 4 × 4 × 4 3D Mesh with four stacked layers connected using TSVs. (c2) Inter-layer connections.

Fig. 2: Floorplan of 2D and 3D BFT topology and Links and TSVs are 128bit wire.

TABLE II: Link length and delay details of 2D and 3D variants of Mesh, BFT.

Topology	Wire(mm)	Delay (clock cycle)	HL (wire) count	VL count	Number of TSVs	Delays(Clock Cycle)
2D Mesh	1.844	4	112	-	-	-
2-layer 3D Mesh	1.844	4	108	32	4096	1
4-layer 3D Mesh	1.844	4	96	48	6144	1
2D BFT	9.376	92	8	-	-	-
	8.976	85	8			
	4.4889	21	16			
	4.088	18	16			
2-layer 3D BFT	8.176	68	8	8	1024	1
	7.776	63	8			
	4.088	18	8			
	3.688	14	8			
	less than 1mm	1	8			
4-layer 3D BFT	4.488	18	8	12	1536	1
	4.088	14	16			
	3.688	16	8			
	less than 1mm	1	4			

IV. Experimental Setup

The cycle-accurate on-chip network simulator is modified to support 2-layer 3D and 4-layer 3D NoC with accurate delay. XY and XYZ routing is used for 2D and 3D Mesh respectively. Horizontal wire delays were modelled as described in the Section III and the TSV delays were modelled from existing works. Delays are estimated based on the floorplan as shown in Figure 1 (a), (b), (c) and 2 (a), (b), (c) to get the accurate performance metric. BFT as shown in Figure 2 is implemented in simulator and the degree of routers in three levels 4, 6, 6 respectively. Nearest common ancestor algorithm used for 64 nodes. Table III shows network configuration parameters of 2D and 3D variants of Mesh and BFT topologies.

TABLE III: Simulated Network Configuration.

BookSim Parameter	Value
Topology	2D Mesh & 2-layer 3D Mesh & 4-layer 3D Mesh & 2D BFT & 2-layer 3D BFT & 4-layer 3D BFT
Network Size	64 Nodes
Switches	28 - 64
Traffic	Uniform Random
Number of VCs	8
VC buffer size	16
Simulation time	10^5 cycles

V. Results and Discussion

A. Average Network Latency

Average network latency obtained from Booksim using default link latencies and from floorplan based link latencies are plotted in Figure 4 for 2D mesh. When we use the floorplan values in simulation, an increase in average network latency from 19% to 43% is observed. To observe the accurate performance of the NoC architecture, floorplan and accurate delay estimation is important before the simulation. The link length varies based on the floorplan, in Mesh, the links are uniform, average network latency is difference is less, but as the length of the link varies, average network latency is affected significantly.

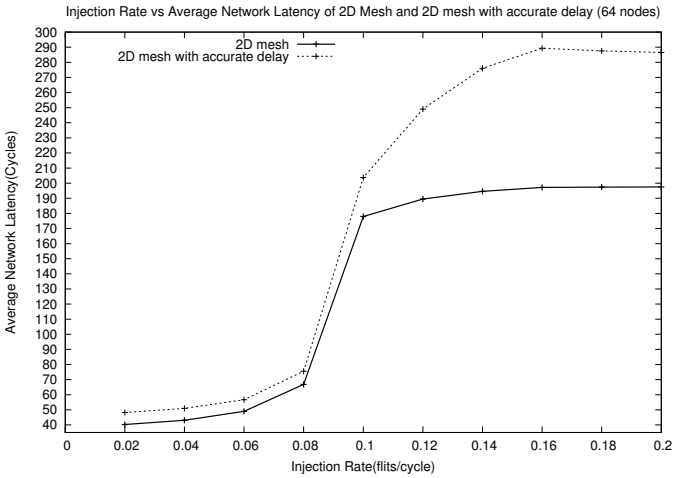


Fig. 4: Average network latency comparison for 2D Mesh and 2D Mesh with accurate delay.

1) 2D Mesh vs 3D Mesh: Figure 5 depicts the comparison of latencies for 2D Mesh, 2-layer 3D Mesh and 4-layer 3D Mesh for uniform with V=8 and D=16. 4-layer 3D Mesh with uniform traffic has reduction in the average network latency up to 54% compare to 2D Mesh for uniform random traffic throughout the simulation.

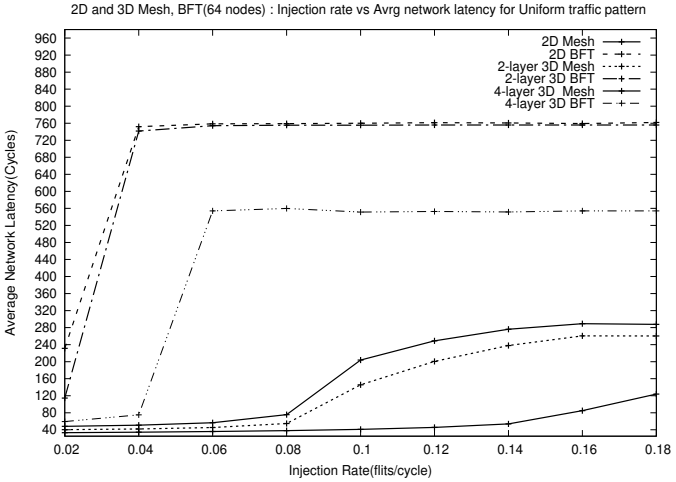


Fig. 5: Average Network latency comparison between all six topology i.e both 2D and 3D variants of Mesh, BFT.

Experiments results show that, the 4-layer 3D Mesh has up to 2.2× performance improvement compared to 2D Mesh for uniform traffic pattern. Based resources usage Table IV, 4-layer 3D Mesh has a reduction in horizontal links from 112 to 96 and there are 48 extra VL compared 2D and because of which there is decrease in latency. In 3D mesh TSV delay is 75% less compare to wire delay and

which lead to decrease in the average latency compare to 2D mesh. In 2-layer and 4-layer 3D mesh, still 4-layer 3D Mesh performance better compare to 2-layer because there are 16 more VL connection compare to 2-layer as shown in resource Table IV.

2) BFT vs Mesh topology: Figure 5 shows the latency comparison of both 2D and 3D variants Mesh and BFT topologies. From Figure 5, 4-layer 3D Mesh has least average network latency compare to all other 2D ad 3D Mesh and BFT topology through out the simulation. 4-layer 3D mesh has 4.5× on-chip communication performance compare to 4-layer 3D BFT. From Table II, link length in Mesh are smaller compare to BFT topology and when we move from 2D to 3D the vertical links delay again reduced thus 4-layer 3D Mesh has least average network latency compare to 3D BFT. There are 36 extra routers and 2× links in Mesh compare to BFT which leads to distributed traffic over the network i.e reduced waiting time so Mesh has least network latency compare to BFT. 2D and 3D BFT has high network latency due to its topology structure and it has least number of resources compare to Mesh and are shown in Table IV.

TABLE IV: Total number of resources used for 2D and 3D variants topology and Links are classified as horizontal links(HL) and vertical links (VL). V is number of virtual channels and D is VC depth per VC

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	8	16	112	0
2-layer 3D Mesh	8	4	2	64	6/6	8	16	108	32
4-layer 3D Mesh	4	4	4	64	6/6, 7/7	8	16	96	48
2-D BFT	4	3	1	28	4/4, 8/8	8	16	48	-
2-layer 3D BFT	4	3	2	28	4/4, 8/8	8	16	40	8
4-layer 3D BFT	4	3	4	28	4/4, 8/8	8	16	36	12

VI. CONCLUSION

We have considered micro-architectural characteristics of on-chip networks such as the floorplan based wire lengths, and link latencies to get more accurate latency values. Wire delays were obtained using ORION delay models and compared the 2D and 3D variants of Mesh and BFT topology by considering uniform traffic pattern through cycle-accurate simulation. 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 0.18 with uniform traffic pattern. 4-layer 3D Mesh has on-chip communication performance upto 4.5× than 4-layer 3D BFT. Among Mesh and BFT, Mesh shows better on-chip communication performance compare to BFT topology and its because of the topology structure i.e links, router input ports, buffers are larger in the 2D and 3D Mesh compared to 2D and 3D BFT. We can vary the number of links and routers in BFT to get optimised design to reach the latency which is near to Mesh topology with less resources. This work considers the floorplan based accurate wire delay and TSVs in 3D variants Mesh and BFT NoCs. Further the thermal issues can be observed to optimise the number stacked layers.

References

[1] D. H. Kim, K. Athikulwongse, S. K. Lim, A study of through-silicon-via impact on the 3d stacked ic layout, in: Proc. of the 2009 International Conference on Computer-Aided Design, ACM, 2009, pp. 674–680.

[2] V. F. Pavlidis, E. G. Friedman, 3d topologies for networks-on-chip, IEEE Trans. Very Large Scale Integr. Syst. 15 (10) (2007) 1081–1090.

[3] J. W. Joyner, P. Zarkesh-Ha, 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.

[4] A. Psathakis, V. Papaefstathiou, N. Chrysos, F. Chaix, E. Vasilakis, D. Pnevmatikatos, 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.

[5] S. Kumar, A. Jantsch, J. P. Soininen, M. Forsell, M. Millberg, J. Oberg, K. Tiensyrja, 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. doi:10.1109/ISVLSI.2002.1016885.

[6] A. Bose, P. Ghosal, S. P. Mohanty, A low latency scalable 3d noc using bft topology with table based uniform routing, in: 2014 IEEE Computer Society Annual Symposium on VLSI, 2014, pp. 136–141.

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

[8] M. Debora, P. Max, K. Marcio, C. Luigi, S. Altamiro, Performance evaluation of hierarchical NoC topologies for stacked 3D ICs, Proc. - IEEE Int. Symp. Circuits Syst. 2015-July (2015) 1961–1964.

[9] A. M. Rahmani, P. Liljeberg, J. Plosila, H. Tenhunen, Lastz: An ultra optimized 3d networks-on-chip architecture, in: 2011 14th Euromicro Conference on Digital System Design, 2011, pp. 173–180.

[10] A. Zia, S. Kannan, H. Jonathan Chao, G. S. Rose, 3d noc for many-core processors, Microelectron. J. 42 (12) (2011) 1380–1390.

[11] A. B. Kahng, B. Li, L.-S. Peh, K. Samadi, Orion 2.0: a fast and accurate noc power and area model for early-stage design space exploration, in: Proc. of the conference on Design, Automation and Test in Europe, European Design and Automation Association, 2009, pp. 423–428.

[12] N. Jiang, D. U. Becker, G. Michelogiannakis, J. Balfour, B. Towles, D. E. Shaw, J.-H. Kim, W. J. Dally, A detailed and flexible cycle-accurate network-on-chip simulator, in: Performance Analysis of Systems and Software (ISPASS), 2013 IEEE International Symposium on, IEEE, 2013, pp. 86–96.

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

[14] P. M. Yaghini, A. Eghbal, S. S. Yazdi, N. Bagherzadeh, M. M. Green, Capacitive and inductive tsv-to-tsv resilient approaches for 3d ics, IEEE Transactions on Computers 65 (3) (2016) 693–705.

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

[16] C. Grecu, P. P. Pande, A. Ivanov, R. Saleh, A scalable communication-centric soc interconnect architecture, in: International Symposium on Signals, Circuits and Systems. Proc., SCS 2003. (Cat. No.03EX720), 2004, pp. 343–348.

[17] G. Katti, M. Stucchi, K. D. Meyer, W. Dehaene, Electrical modeling and characterization of through silicon via for three-dimensional ics, IEEE Transactions on Electron Devices 57 (1) (2010) 256–262.