A Novel 3D Mesh-Based NoC Architecture for Performance Improvement

Navid Habibi[1], Mohammad Reza Salehnamadi[2*], Ahmad Khademzadeh[3]
1- Department of Computer Engineering, South Tehran Branch, Islamic Azad University, Tehran, Iran.
Email: st_n_habibi@azad.ac.ir
2- Department of Computer Engineering, South Tehran Branch, Islamic Azad University, Tehran, Iran.
Email: M_saleh@azad.ac.ir (Corresponding author)
3- Telecommunication Research Center, Tehran, Iran.
Email: zadeh@itrc.ac.ir

Received: July 2021 Revised: October 2021 Accepted: December 2021

ABSTRACT
Applying semiconductor technology, Network-on-Chips (NoCs) are designed on silicon chips to expand on-chip communications. Three-dimensional (3D) mesh-based architecture is also known as a basic NoC architecture characterized by better energy consumption and latency compared with two-dimensional (2D) ones. Recently developed architectures are based on the regular mesh. However, there are serious drawbacks in NoC architectures including high power consumption, energy consumption, and latency. Therefore, improving topology diameter would overcome these shortcomings. Accordingly, a new 3D mesh-based NoC architecture is proposed in the present study utilizing the star node, consisting of a new 3D topology with a small diameter and new deadlock-free routing. The diameter of this architecture is then compared with its counterparts. Afterward, the scalable universal matrix multiplication algorithm (SUMMA) is implemented in the proposed architecture. The results indicate a smaller network diameter, lower energy consumption (32%), less network latency (8.6%), as well as enhancement in throughput average (13.6%). The proposed matrix multiplication algorithm also implies improvement in the cost of the proposed architecture in comparison with its counterparts.

Keywords: Communication, Network Architecture, Topology, Network-on-Chip, System-on-Chip, Routing Protocols, De Bruijn Graph, Performance Evaluation, Multiplication Algorithm, Latency, Energy Improvement, Diameter.

1. INTRODUCTION

Based on Moore's law, the number of transistors doubles about every two years, and on-chip components are also on the increase. Likewise, communication is considered a crucial issue in the performance of a system on a chip (SoC). Accordingly, network-on-chip (NoC) refers to a solution for long links in this circuit [1], making communication in such networks is a significant problem. Additionally, NoC represents the concept of an integrated micro-network on a silicon chip wherein each core consists of intellectual property (IP) core, an interface, and a router, which provides higher bandwidth and reduces latency [2]. Over recent years, the energy consumption of NoCs has augmented, leading to the generation of higher levels of power and heat in these networks. A proper NoC architecture has thus adequate power, efficient energy consumption, and good latency [3].

The performance of NoCs depends on some properties like topology, routing algorithm, selection strategies, as well as switching and mapping techniques [4]. Via expanding two-dimensional (2D) topologies into three-dimensional (3D) ones, the advantages of lower cost and higher performance can be assessed. The 3D NoCs can be additionally expanded and applied vastly in modern technologies. Such networks are similarly utilized in different areas with various metrics mentioned for NoCs [5].

The extra dimension in NoCs also needs bigger routers with more ports. Some 2D planes are accordingly connected by vertical links of the 3D NoC architecture and traditional NoC routers have five ports. The 3D architectures with lower connectivity length provide higher connectivity of nodes. As well, it consumes less energy and power while having better latency compared with traditional 2D architectures [6]. Routers in 3D NoC have seven ports: two up and down ports for 3D

Paper type: Research paper
DOI: https:// 10.30486/mjee.2022.696497
How to cite this paper: N. Habibi, M.R. Salehnamadi and A. Khademzadeh,"**A Novel 3D Mesh-Based NoC Architecture for Performance Improvement**", Majlesi Journal of Electrical Engineering, Vol. 16, No. 2, pp. 85-101, 2022.

connection, one port for connecting local IP-core, and four cases for 2D connections [7].

Matrix multiplication is the most common mathematical operation used in many fields such as physics and computers. Performance of the multiplication has been correspondingly improved by many researchers in $O(n^3)$ time in a 2D mesh-based architecture to $O(\log n)$ in 2-by-2 parallel random-access machine (PRAM) mesh-based architecture [8].

Interconnections of the NoC use about 65-80% of its total power consumption [9]. In this study, a novel 3D NoC architecture comprised of a new topology with a small diameter and a new routing algorithm based on this deadlock-free topology is proposed. By decreasing hop count from each source node to destination one, this proposed architecture would have less latency, lower energy consumption, and better performance than its counterparts.

The organization of this paper is as follows: the literature review is presented in Sec. 2, the De-Bruijn graph is described in Sec. 3, the scalable universal matrix multiplication algorithm (SUMMA) is defined in Sec. 4, the newly designed 3D NoC architecture is proposed in Sec.5, performance evaluation is run in Sec. 6, and the study is concluded in Sec. 7.

2. RELATED WORKS

Different topologies have been proposed in 3D NoC architectures. Various NoC architectures have been also proposed in [1]. The 3D mesh-based architecture (Fig. 1-a) is an $m \times n$ mesh of IP-cores whose switches are interconnected. Its routers also have seven ports, two of which are connected to up and down horizontal planes of the IP-core. As well, it is a regular mesh structure, applied in 3D architectures. Different type of routers and large diameter leads to the stacked mesh (Fig. 1-b) which is similarly a combination of 2D mesh layers, wherein one bus connection is applied in vertical connections. Limitation in bus connections is its disadvantages besides router complexity. Considering the ciliated mesh (Fig. 1-c), there is only one 2D mesh. Other layers are directly connected to the switches in the first main 2D horizontal plane. This architecture has a small diameter but, routers would have congestion. The 3D butterfly fat tree (BFT) (Fig. 1-d) also resembles the 2D BFT in a mesh structure with the limited number of connections. Its disadvantage is long link lengths which makes extra latency. Among these four architectures, the ciliated mesh reveals better energy consumption in comparison with its counterparts. A new 3D NoC architecture based on the De-Bruijn graph has been correspondingly proposed in [10], in which a switch named enhanced pillar structure is used to connect all layers. A routing based on shifting address which is deadlock-free and has fault tolerance is additionally utilized. This architecture is named 3D DB (Fig. 1-e), which reduces diameter only in every 2D horizontal plane. Normal connections between 2D layers make long latency in networks which have planes more than five. In the 3D De-Bruijn architecture, no difference is observed in horizontal planes and connection construction, which is assumed as a limitation of this architecture. Whenever the third dimension of the topology is large, latency in this network becomes of concern.

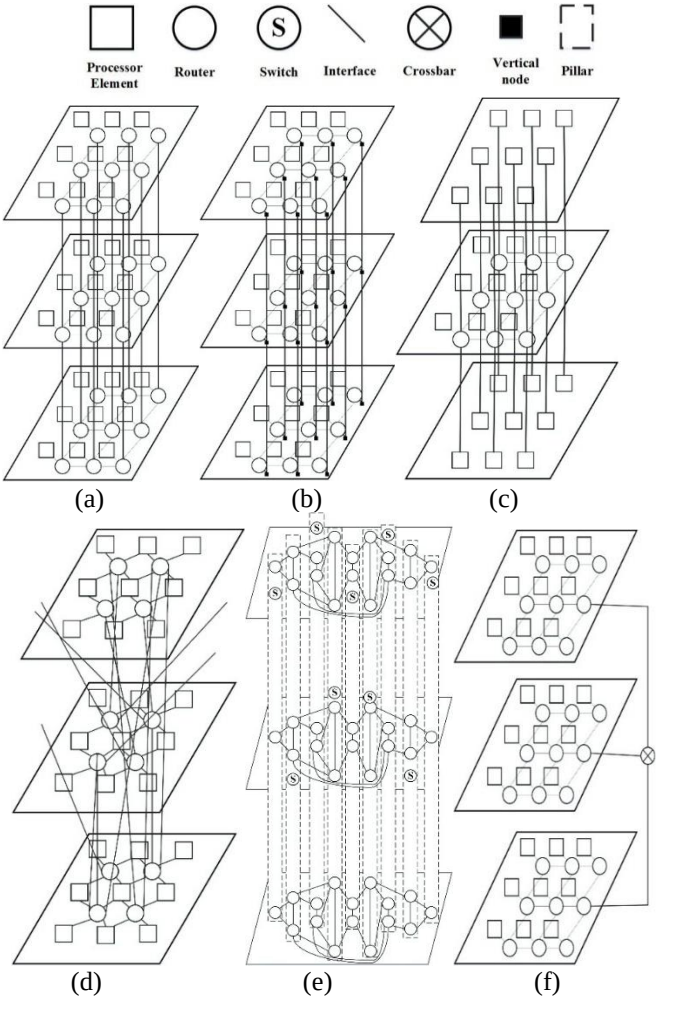


Fig. 1. 3D NOC architecture designs (a) Normal mesh (b) Stacked mesh (c) Ciliated mesh (d) 3D BFT (e) 3D DB f. DB_EP [1, 10 and 11].

A crossbar switch has been further applied in [11] to connect the 2D horizontal planes of IP-cores, named DB_EP (Fig. 1-f). Using a crossbar switch to connect a single IP-core from each one of the horizontal planes also causes bottlenecks; thus, a limitation arises in the number of the horizontal planes. Moreover, long-length links make a big latency in the network. In networks with more than 5 planes, long links lengths, switch deadlock and congestion would occur.

A new NOC hierarchical architecture for mapping neural networks to chips is introduced in [12]. It also concludes an optimization of energy consumption and network Latency in NOC communications. A short communication NOC architecture is applied in [13]. This architecture is an optimized area and power of NOC with improvement the architecture if the router and decreasing network diameter with shortcoming source to a destination hop count.

A new 3D recursive network topology, as a mesh-based one with a cluster head in each cluster has been additionally proposed in [14]. To connect the vertical links, through silicon via (TSV) is applied where each layer has four nodes containing a cluster head. Each IP-core also has a three-digit address and its routing is based on a recursive address. Expanding the network in all dimensions is a limitation in this architecture. A new 3D architecture based on the dense graph has been introduced in [15], in which a dense Gaussian structure is used on a NoC to obtain a small diameter with a new all-to-all broadcast routing. This architecture has large complexity and bigger link length than traditional mesh architectures. Analyze of expansible NoC architecture and a comparison of the related works on that are presented in [16].

In the smart model, a smart router can bypass flits to pass through other routers, with help of some multiplexers and repeaters. An analytical model of Smart NoC is proposed in [17] that predicts latency and throughput of smart NoCs and reduces simulation time. A smart model is proposed in [18] that virtually bypass all routers of the packet route within a single cycle, without adding a physical channel. This model uses asynchronous repeaters and circuit switch repeaters and reduces simulation time.

A 2D hexagonal mesh with a significant routing has been employed in [7] and then compared with a 3D mesh. This architecture consumes lower power than 3D mesh with XYZ routing while it uses long link lengths. Degrees of freedom in routers are all 6 with an extra diagonal routing. A heterogeneous honeycomb topology with a complicated deterministic routing has been further applied in [19] which costs 20% lower than mesh architecture and can moderate delay in comparison with mesh architecture.

An enhanced dynamic XY (EDXY) algorithm has been proposed in [20] and then implemented in a 2D mesh topology. This architecture shows improvement in latency in comparison with 2D mesh architecture with XY routing while it is not expanded in 3D design. Recursive network topology with recursive address routing has been additionally applied in [21]. A new router is proposed in [22] and the size of the buffer is mentioned. Energy consumption and quality of the service are improved. A new NoC design is proposed in [23] which improves latency with balance in tradeoff with latency power consumption and chip area which has an increase in area.

To provide an application, Parallel matrix multiplication algorithms have been described and compared in [8]. There are some multiplication algorithms run in 2D and 3D NoC mesh architectures, such as SUMMA, Fox, and Cannon which are known as the latest ones. Among the given algorithms, the SUMMA is proper in terms of the 3D mesh in NoC architectures. The 2D Cannon matrix multiplication algorithm is also concerned with an n-by-n mesh architecture. This algorithm reduces multiplication steps in a 2D multiplication [24]. In this paper, a multiplier algorithm is provided to illustrate the proposed architectural application.

3. THE DE-BRUIJN GRAPH

The De-Bruijn graph is a DB (d,k) with N nodes and degree 2d which has a diameter of $\log_d N$ and total nodes of $N = d^k$ [25]. Some implementations of this topology in the industry have been thus far expressed in [26, 27, 28, and 29]. Besides, this graph is applied in many NoC architectures. Therefore, two nodes connect if one of the Eqs. (1) or (2) holds true:

$$i = (d \times j + r) \bmod N, \quad r = 0,1,\ldots,d-1 \tag{1}$$

$$j = (d \times i + r) \bmod N, \quad r = 0,1,\ldots,d-1 \tag{2}$$

The De-Bruijn is a k bit d digit array that changes states by shifting address. There is also a node with an identifier $(i_{k-1},\ ik_2,\ \ldots,\ i_1,\ i_0)$, where $i_j(0,\ 1,\ \ldots,\ d-1),\ 0 < j < k-1$ and its neighbors are $(i_{k-2} \ldots i_1 i_0 p)$ and $(p i_{k-1} i_{k-2} \ldots i_1)$ in which $p(0,\ 1,\ \ldots,\ d-1)$ [11, 28]. The De-Bruijn graph is shown in Fig. 2.

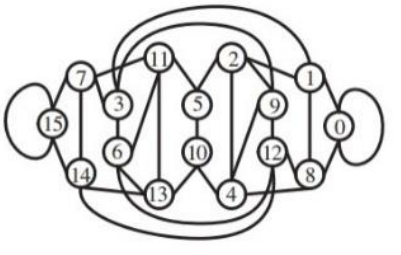


Fig. 2. DB (2,4) topology [11].

The De-Bruijn topology has smaller diameter compared with other topologies. Using this structure in the NoC architecture helps reducing the diameter of one part of the architecture with a fixed structure. Diameter of De-Bruijn topology is logarithmic and is used in the proposed architecture.

4. SUMMA ALGORITHM

Assume $C = A \times B$, where C is a result of multiplying $A = \{A_{ik}\}$ and $B = \{B_{kj}\}$ in n-by-n matrix, $1 \le i, j, k \le n$. In a NoC with a P processor in the form of $X \times Y \times Z$, Z is layers of the architecture and each layer has $X \times Y$ mesh. Z dimension is also used to parallel the multiplication of the matrices and to diminish computation overload through expanding communication of processors. The idea of 3D SUMMA is given by expanding Cannon and Fox ones [8, 30]. The SUMMA thus represents less cost than other multiplication algorithms [31].

Partitioning A and B in $\frac{k}{Z}$ column and row, $A = (A_0 \mid \ldots \mid A_{z-1})$ and $B = \begin{pmatrix} \frac{B_0}{B_1} \\ \frac{\ldots}{B_{z-1}} \end{pmatrix}$. Then, the result is as follows:

$$C = C + AB = (C_0 + A_0 B_0) + (C_1 + A_1 B_1) + \ldots + (C_{z-1} + A_{z-1} B_{z-1}) \quad (3)$$

$(C_0 + A_0 B_0)$ is calculated in layer zero of the 3D mesh-based NoC architecture. $(C_1 + A_1 B_1)$ is accordingly calculated in layer one of the mesh architecture in parallel algorithm and $(C_{z-1} + A_{z-1} B_{z-1})$ is calculated in the last layer, z-1 [8, 30]. The SUMMA is presented as follows [31]:

1. Algorithm: *C* := A(*C; A; B*)
2. Partition *C*→*(CL | CR)* , *B*→*(BL | BR)*
3. where *CL* and *BL* have 0 columns
4. While *n*(*CL*) < *n*(*C*) do
5. Determine block size *b*
6. Repartition
7. *(CL |CR)*→*(C*0 |*C*1 |*C*2), *(BL |BR)*→*(B*0 |*B*1| *B*2)
8. where *C*1 and *B*1 have *b* columns
9. *B*1(∗*; MR*) ←*B*1(*MC; MR*)
10. *C*(*t*) 1 (*MCs ;* ∗) := *A*(*MCs;MRt*)*B*1(*MRt ;* ∗)
11. *C*1(*MC; MR*) := $\sum tC1(t)(MC; *)$
12. Continue with
13. *(CL |CR)*→*(C*0 |*C*1 |*C*2), *(BL |BR)*→*(B*0 |*B*1| *B*2)
14. End While

Where in that [31].

RELATION	DEFINITION
$bj(VC;*)bj(MC;\ MR)$	Scatters within rows
$bj(VR;*)bj(VC;*)$	Permutation
$bj(MR;*)bj(VR;*)$	All gathers within cols
$cj(MC;*)\ := A(MC;\ MR)$ $bj(MR;*)$	Local matrix-vector multiplications
$cj(MC;MR) \leftarrow \sum cj(MC;*)$	Reduce-to-one within rows

This algorithm runs in the following steps [31]:

I. Matrix C is partitioned and each processor holds a partition in $C(CL \mid CR)$, $B(BL \mid BR)$ step.

II. In the scatter plot with rows and columns, each processor holds a column of matrix A and a row of matrix B in $B1\ (*;\ MR)B1(MC;\ MR)$ step.

III. Multiplication is performed in each processor in $C(t)1\ (MCs;*) := A(MCs;\ MRt)\ B1\ (MRt;*)$ which is a permutation step.

IV. Each processor updates its block of matrix *C* in $C1\ (MC;\ MR) := \sum tC1(t)(MC;\ *)$ which all gather within columns and then reduce to one within row step.

V. Blocks update and then the algorithm runs once again in $(CL \mid CR)(C0 \mid C1 \mid C2)$, $(BL \mid BR)(B0 \mid B1 \mid B2)$ step.

5. PROPOSED 3D NoC ARCHITECTURE

This 3D NoC architecture is made up of components such as topology, routing, switching, and flow control, among which topology and routing constitute the two determinative main parts in network performance. A proper topology with a smaller diameter would thus contribute to less latency and lower energy consumption throughout the links, in addition to network path diversity. A proper topology is thus efficient if and only if its routing is so. A proper routing that conducts packets to a route in an appropriate path is also of importance and it is defined in this topology. This routing should be deadlock-free to avoid deadlock or high latency in the network.

5.1. Topology

The proposed topology consists of some elements as presented and defined:

Definition 1: Vertical connections (VCs) are vertically connected to routers in horizontal planes in a 3D NoC architecture. These routers are also connected to IP-cores. To connect these routers, the star-router (Rs) is applied in NoC horizontal planes.

Definition 2: Horizontal planes (HPs) are wafer layers designed and implemented on a NoC. These planes are connected through the VCs.

Definition 3: Star-routers (Rs) are routers with different connections applied in decreasing network diameter in this proposed architecture.

Each horizontal plane is constructed from some nodes, connected to their neighbors. In this proposed 3D topology, there are two types of nodes: the R nodes, i.e. typical routers with their regular connection, and the Rs

nodes (i.e. star nodes). These nodes are applied with different connections in the topology. By applying some of these nodes between layers, the dynamic routing instead of static one would yield. Each Rs node can even transmit some nodes from the top layer to the bottom one. These nodes are applied to assist packet transfer between the layers. One Rs node with its connections is shown in Fig. (3), where a switch star node is a central node connecting three nodes of the upper layer to two other nodes of Rs horizontal plane. Moreover, all Rs nodes of HP (i-1), HP (i), and HP (i+1) are connected. As observed here, one of these proposed network nodes is applied in the layers to accomplish this connection. These Rs nodes also transmit upper layer R nodes to that of layer R nodes. In this study, the De-Bruijn graph is utilized to decrease latency in the second *Y* dimension.

In this process, each node has two options in terms of being connected to other horizontal plane nodes, namely, by regular vertical link and by going to another horizontal plane node. The decision is also made when the node goes to the other half of the network. The movement can be thus normal at its vertical link to connect to R node or diagonal to connect to Rs node. Once one node wants to go to other layers, it can be present at Rs node and decide which node has smaller hop counts towards the destination node and this happens in just one vertical hop count.

An Rs node connects two nodes in the upper layer (layer *k+1*) to the layer *k* nodes. All central nodes are also connected to one another and act as an elevator. In this situation, some nodes in three layers are directly connected to a switch star node, that is, these nodes are connected to one another with just two hop counts. In the worst case (Fig. 3), the distance from the source node to destination one is three hop counts at its maximum. In regular mesh connections in 3D NoCs, the distance would be four hop counts. It is because of the Rs node which provides the opportunity to choose and to be connected to another horizontal plane node in fewer hop counts.

This proposed topology is divided into three main parts. Assume $X \times Y \times Z$ as the topology dimensions. The first part is the construction of the first *X* dimension (Fig. 4-d). Among all nodes in this dimension, there is an Rs node. These nodes connect each horizontal plane to other upper and lower horizontal ones. The second part is the construction of the first *Z* dimension. The main idea of this proposed topology is in the construction that decreases the network diameter. The construction of the third dimension based on the Rs node also shortens the path between the source and destination nodes. By moving in the third dimension connections, *X* dimension can shorten the path for packets. The diagonal movement of Rs node similarly contributes to having fewer hop counts in the first dimension when the third one moves. The third part is the construction of the first *Y* dimension which is correspondingly constructed through the De-Bruijn graph with a small diameter and the same number of connections towards mesh construction contributing to the topology to have a low latency in this dimension.

The main idea of the topology is on the third dimension connections, applied through different vertical links by a star node. An example of $5 \times 4 \times n$ topology with *n* horizontal plane is shown in Fig. 4, where in section (a) an Rs connection is observed in this topology.

This switch connection includes connections of Rs nodes up and down layer, connections of its row nodes, and connections of two other upper layer nodes. This Rs node also connects all six nodes with one another. All Rs connections are illustrated in Fig. 4-b and the normal vertical connections are shown in Fig. 4-c.

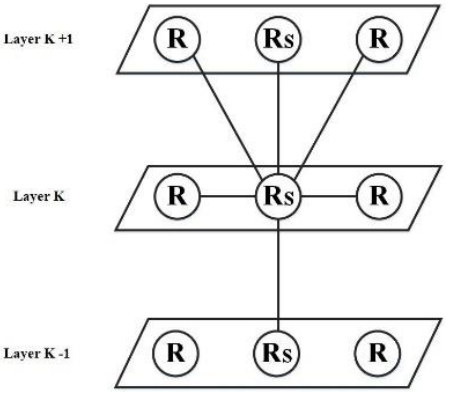


Fig. 3. One Rs Node (star node) Connections in this proposed 3D Architecture.

With regard to $X \times Y \times Z$ topology, Figs. 4-d and 4-e represent its *X* and *Y* connections; respectively. Connections in the first dimension (*X*) are based on the Rs node and the third dimension construction. The second dimension (*Y*) is also constructed with reference to the De-Bruijn graph, which contributes to having less latency and lower energy consumption with a smaller diameter. All together, these connections construct a topology to shape this proposed 3D topology.

Connection conditions of the proposed Topology is as follow:

a. First Dimension:

$$CL(i,j,k) = (i \pm 1, j, k) => \begin{cases} i = i \pm 1 \\ j = j \\ k = k \end{cases}$$

b. Second Dimension:

$$CL(i,j,k) = (i, 2i \bmod n, k) => \begin{cases} i = i \\ j = 2i \bmod n \\ k = k \end{cases}$$

c. Third Dimension:

I. in normal nodes

$$CL(i,j,k)=(i,j,k\pm1)=>\begin{cases} i=i \\ j=j \\ k=k\pm1 \end{cases}$$

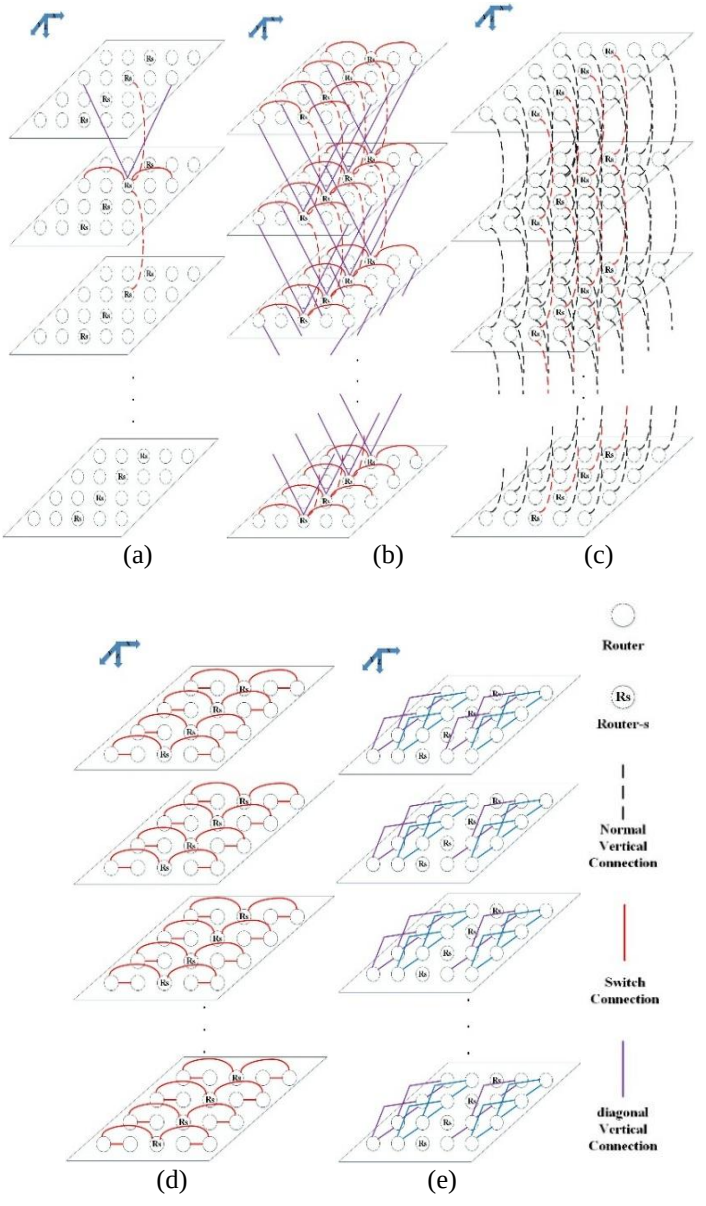


Fig. 4. Example of a 5 × 4 × n in the Proposed Topology Separated in Each Dimension (a) One sample for Rs connections (b) Whole Rs connections (c) Normal up/down port connections (d) First dimension (x) connections based on Rs e. Second dimension De-Bruijn connections (y) connections

II. in Rs nodes

$$CLh(i,j,k)=\left(i,\sum_{y=\frac{j}{n}}^{\frac{j}{n}+j}y,k+1\right)=>\begin{cases} i=i \\ j=\sum_{y=\frac{j}{n}}^{\frac{j}{n}+j}y \\ k=k+1 \end{cases}$$

Where the cluster Condition is as

$$N_{cl,2D}=\frac{j}{n},N_{clh,3D}=k\times\frac{j}{n},N_{cl,3D}=k\times\frac{j}{n}.$$

$N_{cl,2D}$ Presents clusters in a 2D dimension, $N_{cl,3D}$ presents cluster in 3D dimension and $N_{clh,3D}$ is cluster-head count in 3D dimension.

The connections of the third dimensions include vertical normal, Rs, and X dimension connections, in a $5 \times 4 \times n$ topology as an example.

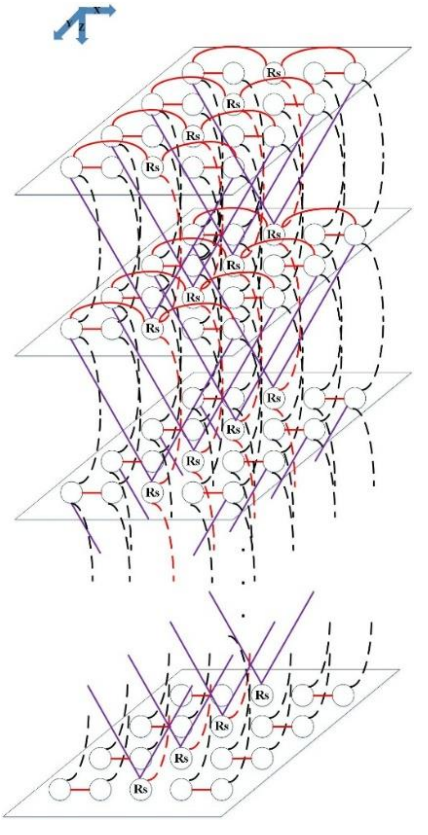


Fig. 5. Example of a 5 × 4 × n with n horizontal plane in this Proposed Topology, connections in all three dimension

5.2. Routing

The proposed routing algorithm is deadlock-free and applies the third dimension to move in the shortest path. For the routing algorithm, the priority is on moving dimensions. In the priority, packets move in the third dimension, wherein the distance of the source node to the destination one is calculated as Δz. If Δz is one, the

movement is on the shortcut path and the network packets move to the shortcut path; otherwise, the movement is on the normal path in $\Delta z - 1$ step. To have a shortcut path, its probability should be assured, that is the position of the source and destination nodes should be checked. Then, packets decide whether the shortcut path is needed or not. The proposed routing is divided into two parts:

Part one is to assure that both source and destination nodes are in the same half (both in the left or right half of *X* dimension). In this situation, the destination node is closer to the source node, and the shortcut path is not required.

Part two is to confirm that both source and destination nodes are not on the same side. In this condition, a shortcut path is chosen to jump to the nearest destination. In the worst case, two hop counts are necessary in the *X* dimension, indicating that the diameter of the network is changed without adding a long connection.

This deadlock-free routing algorithm for this proposed 3D topology has two steps which are considered as decisions on moving in the third dimension expressed as follows:

Step 1: Determining whether the destination node is in the same horizontal plane or not, if yes, routing is needed to be done in two dimensions only.

Step 2: If the destination node is not in the same plane, there is a need to route using vertical connections and Rs. Based on the decision made in the first step, packets move through the third dimension in the most appropriate path. This allows packets to have fewer hop counts in the *X* dimension when they cross *Z* dimension. In the second step, the first dimension is initially passed in fewer hop counts, followed by the second dimension routing.

The proposed routing is deadlock-free. Deadlock also occurs when a group of agents in a network waits for each other. This arises in a waiting cycle graph with resource dependency, related to the applied routing algorithm. If the routing algorithm in the network is free of the cycle graph, that network uses a deadlock-free routing, because there would be no waiting cycle for the agents. Restricting a dimension in the proposed routing can also omit the cycle in the routing algorithm. First, the third dimension is routed completely, followed by routing the first dimension, and then, the second one. As the third dimension is restricted, there will not be a dependency cycle in the third dimension. Through restricting routing in a horizontal plane in the first and second dimensions, there will not be a dependency cycle. In accordance with this proposed ordering routing algorithm, the deadlock never occurs in this 3D topology, due to restricting the routing. The flowchart of the proposed routing is shown in Fig. 6 where in line 22, there is log(y) move based on De-Bruijn topology that is used in this dimension.

Proposed Routing Pseudo Code

```
1: function Proposed-Routing
2: dst_coord = destination coordination;
3: pos = position;
4: position.z - dst_coord.z = Δz;
5: if (dst_coord.z > position.z)
6:     if (Δz != 1) go to UP PORT
7:     else if (pos.x && dst_coord.x in same half of
x Dimension) go to UP PORT Then goto line 4;
8:     else if (pos.x right half of x dimension &&
dst_coord.x in left half of x Dimension) go to WEST
PORT;
9:     else if (pos.x left half of x dimension &&
dst_coord.x in right half of x Dimension) go to EAST
PORT;
10:    end if
11:    end if
12: if (dst_coord.z < position.z)
13:    if (Δz != 1) go to UP PORT
14:    else if (pos.x && dst_coord.x in same half of
x Dimension) go to DOWN PORT Then goto line 4;
15:    else if (pos.x right half of x dimension &&
dst_coord.x in left half of x Dimension) go to EAST
PORT;
16:    else if (pos.x left half of x dimension &&
dst_coord.x in right half of x Dimension) go to WEST
PORT;
17:    end if
18: end if
19: Xi=Destination.X – Source.X;
20: if (xi > 0) go to WEST PORT;
21: else go to EAST PORT;
22: Move Log(y) to destination Y;
23: end function
```

6. RESULTS AND DISCUSSION

The proposed architecture in this study evaluates latency and energy consumption along with network diameter, SUMMA cost on the proposed architecture, and throughput. This new proposed 3D NoC architecture is simulated based on the cycle-based simulator. The Noxim simulator is also applied as open source software, this architecture is simulated, and then average energy, packet latency, and maximum network latency are calculated. The following simulation results are on the basis of the proposed routing algorithm, topology, random selection strategy, and random traffic type with 10000-cycle simulation time. The exact configurations for the simulation of the proposed architecture are presented in Table 1. The results of the simulations describe latency, average latency, total energy consumption, and throughput.

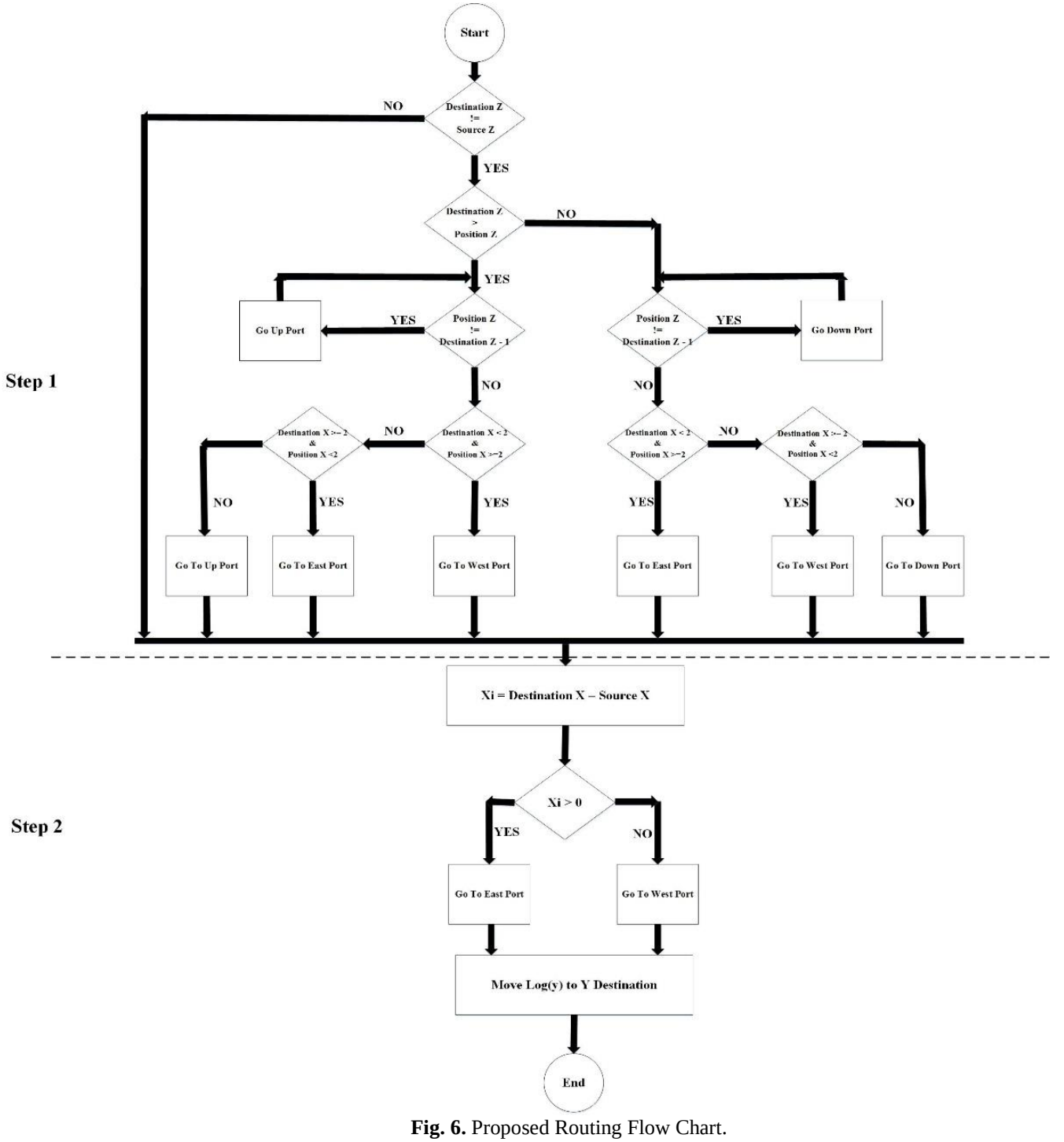


Fig. 6. Proposed Routing Flow Chart.

Table 1. Network On chip Configurations

Name	Value
TOPOLOGY	Proposed 3D Topology
Dimension	Different Sizes
Ports	six
Routing	Proposed Routing
Selection Strategy	Random
Traffic Type	Random
Simulation Time	10000 Cycle
buffer depth	Four flit
warmup time	1000
injection rate	0.01 – 1 (Poisson)
buffer size	4 flit

6.1. Average Network Latency

Average network latency refers to the average time a packet needs to pass from the source node to the destination one. In Noxim simulations, different packet latencies are also obtained by considering this proposed 3D architecture in various dimensions like $5 \times 4 \times 4$, $5 \times 3 \times 6$, and others. These latencies are presented in different packet injection rates. Moreover, the results of simulations indicate better packet latency which can be additionally calculated through Eqs. (4) and (5) [32]:

$$Latency = T_h + T_s \quad (4)$$

$$T_s = \frac{L}{b}, T_h = H.t_r + H.t_w \quad (5)$$

Where, T_h is time for header flits to pass through a link between two neighbor nodes and T_r and T_w refer to time for a bit to pass a router and a wire; respectively. To calculate the whole network latency, an H hop count is multiplied. Moreover, T_s shows serialization time, L represents link length, and b denotes the bandwidth of that link.

Theorem 1: In a direct relation in symmetric NOC, with a decrease in network diameter, Latency decreases (Diameter $\propto Latency$)

Proof: Let Eq. (5) be the latency in a router and a Link for a packet with header and body for H hop count. Eq. (4) is the total Latency for a Packet through the route that gives the expressing of (H $\propto Latency$).

Diameter is defined as a maximum of minimum hop counts between each two pairs of source and destination nodes in the whole network. Thus, (Diameter $\propto H$). Hence, Network Diameter is in direct relation with Network Latency (Diameter $\propto Latency$) as desired.

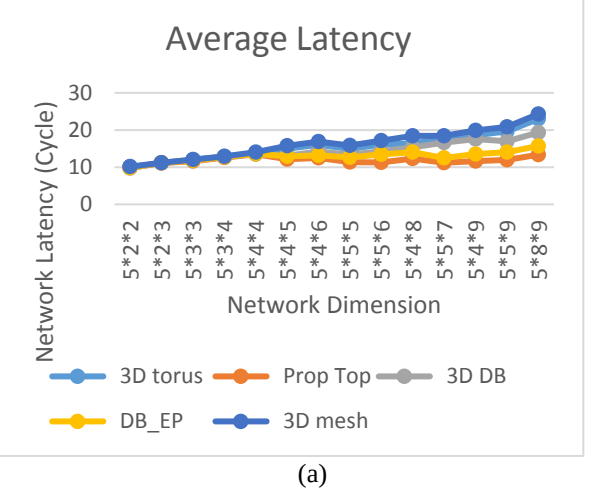
In the proposed architecture, packet length, one hop length, router structure and bandwidth are the same as mesh architecture. Therefore, time takes a packet to pass through a link and a router (T_r and T_w) is the same. Thus, with decrease in h as hop count, latency decreases as well, Eqs. (4) and (5).

This architecture has a smaller diameter than other NoC ones. Diameter is the maximum of minimum hop counts among all two pairs of source and destination nodes, that is, smaller diameter means smaller hop counts between each source and destination node. A comparison of diameter in the latest NoC architectures is listed in Table 6. So:

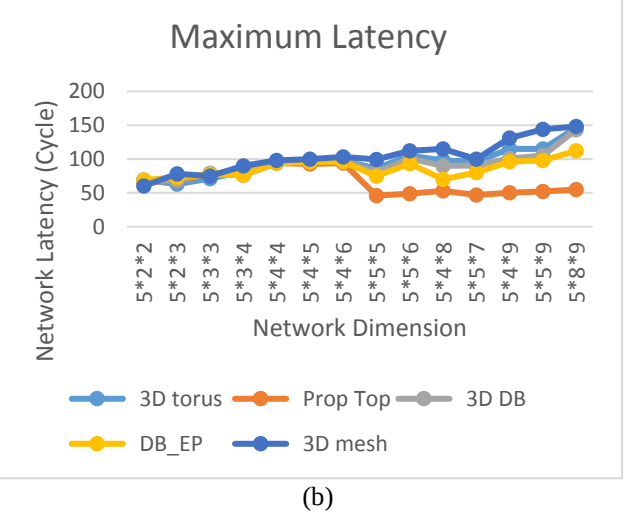
$$Destination \rightarrow Source = H \times (one\ hop\ count\ latency)$$

Small diameter means small hop counts; thus, it is as a smaller H. The simulation results show less latency for packets that transmit the source node to the destination one. A better latency across other architectures is shown in Fig. 7, where the red line represents this proposed architecture. Once the number of nodes increases in more than a hundred nodes in $5 \times 4 \times 4$ dimension, the latency through different dimensions decreases. All these differences are due to this new 3D architecture. This proposed architecture as well as different connections of the third dimension is similarly revealed when nodes are added, and the third dimension grows where packets transmit through the source to the destination route in a very shorter time in other architectures. As observed in Fig. (7), in more than a hundred nodes, the maximum latency and average latency augment with a significant difference.

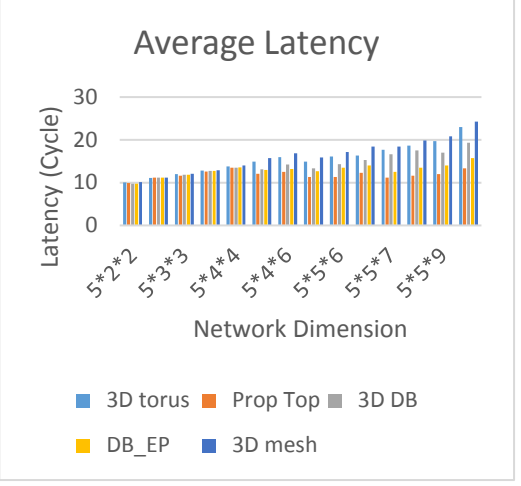
The average network latencies in different architectures are compared in Fig. 7-a. The differences are clear in Fig. 7-c where the bars show better latency in various dimensions. As the network grows, latency in different architectures also increases. The maximum latency of the simulated architectures is illustrated in Fig. 7-b and the bar chart of the maximum latency is shown in Fig. 7-c, revealing a less network latency in the worst case in the network.



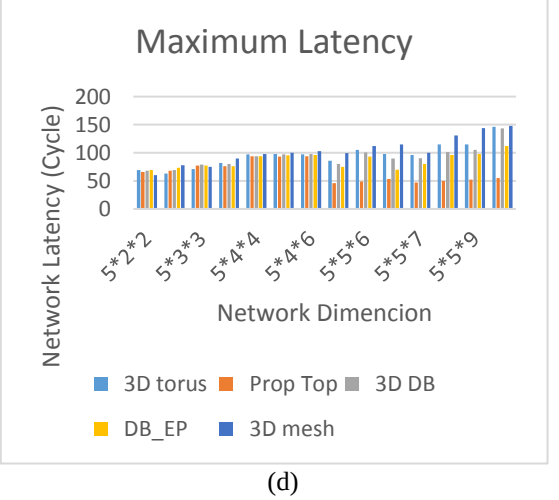
(a)



(b)



(c)



(d)

Fig. 7. (a) Line Chart of Average Latency and (b) Line Chart of Max Latency of Different Architectures (c) Bar Chart of Average Latency and (d) Bar Chart of Max Latency of Different Architectures.

With the increase in network dimension, size of networks will increase. In direct relation, with an increase in network size, average network latency and maximum network latency increase, Fig. 7. In each network dimension, the proposed architecture shows an improvement in latency than its counterparts, Fig. 7.

In networks with a small number of nodes, the distance between every two pairs of source and destination nodes is small, making the differences in latency of these networks non-significant. There is equally a direct relationship between network growth and difference in latency. In large networks, the difference in diameter is much bigger than small NoCs. So, latency in this proposed network is much less and even different. Latency in various dimensions of architecture is depicted in Fig. 7. Latency differences in various injection rates in a $6 \times 6 \times 10$ network dimension is presented in Fig. 8. In the bar chart and the line chart in Figs. 8-a and 8-b, an average 8.6% improvement is shown. With increase in network injection rate, network latency increase. In injection rate of 0.3 flit/node/cycle, network saturates and thus, latency remains constant in average of more than 5600 cycle. The proposed architecture's latency reveals 8.6% improvement toward its counterparts, Fig. 8.

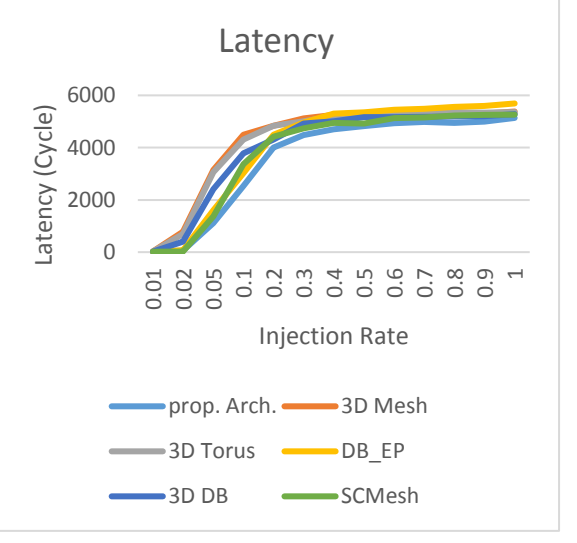
Table 2. Average Network Latency (Cycle).

Architecture	Injection Rate 0.05
3D Mesh	*3145.68*
3D Torus	*3062.25*
3D DBG	*2414.84*
3D DB_EP	*1601.82*
3D SCMesh	*1372.15*
Proposed Arch. (3 nodes in cluster head)	*1116.3*

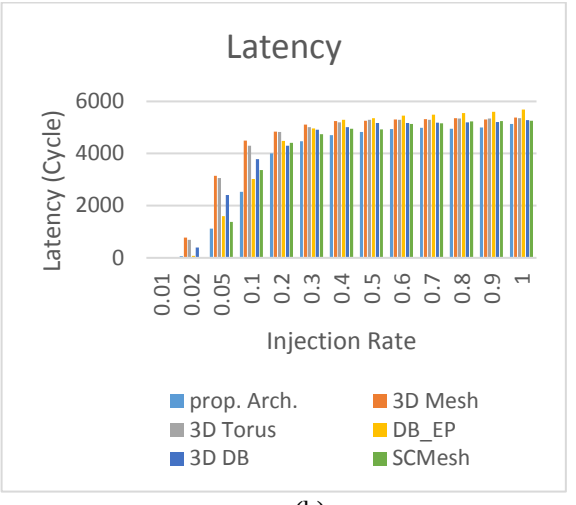
Table 2 outlines the average node latency in PIR of 0.05 in different architectures. It is obvious that average latency increases as network PIR increases and average network latency decreases as NoC architecture is improved.

6.2. Energy Consumption

Due to the limitations of resources in nature, applying some techniques to optimize NoC operations is of the essence [3]. A large number of these techniques include a new topology with an optimized diameter and a routing technique. This proposed topology has a small diameter and its proposed routing algorithm is deadlock-free with path diversity.



(a)



(b)

Fig. 8. a. Line Chart of Network Latency in different network injection rates b. Bart Chart of Network Latency in different network injection rates

These techniques allow network on chips have lower energy consumption. As one of the important metrics of the designs, energy is calculated through Eq. (6) and (7), [1, 9, and 33]:

$$E_{bit} = h \times ERbit_{sw} + h \times ELbit_{link} + 2 \times ECbit_{link} \quad (6)$$

$$E_{NOC} = k \times \left[h \times ERbit_{sw} + h \times ELbit_{link} + 2 \times ECbit_{link} \right] \quad (7)$$

Where, E_{bit} energy of one bit is necessary for it to be transmitted from the source node to the destination one, and *h* refers to hop count. Each movement of a link from a node to its neighbor is a hop count. $ERbit_{sw}$ also shows energy of a bit passing a router. As well, $h \times ELbit_{link}$ denotes energy required by a bit to pass a link between two nodes in a network and $ECbit_{link}$ is energy consumed between router and links. To consume E_{NOC} as the total energy used in this NoC, the energy of one bit is multiplied by total k-bit that should be transmitted.

Eqs. (6) and (7) reveal energy consumption in NoC, in two parts of switch and links, allowing the calculation of energy for each bit in NoC. The energy consumption of the bit is for the links and switches. In *h* hope count, this energy is required. To calculate the accurate energy consumed for each bit, energy consumption of the source and destination nodes is also added to Eq. (6) for k-bit and this relation is multiplied by *k*.

Theorem 2: In a direct relation in symmetric NOC, with decrease in network diameter, Energy consumption decrease (Diameter$\propto Energy$)

Proof: Let Eq. (6) be the Energy consumes for a bit in the whole network in switches and links in H hop count. Eq. (7) is the total Energy consumes in the network of a packet with k bit length. Eq. (7) gives the equation expressing (H $\propto Energy$) as desired.

Diameter is defined as maximum of minimum hop counts between each two pairs of source and destination nodes in the whole network. Thus, (Diameter $\propto H$). Hence, Network Diameter is in a direct relation with Energy consumption in a Network (Diameter $\propto Energy$) as desired.

In the proposed architecture, Wire length, Packet size, size of the capacitor, node's count and switch structure are all same as mesh architecture structure. Thus in Eq. (6), energy of a separate link and router is the same as mesh, due to proposed architecture. Therefore, with decrease in h parameter in Eq. (6), energy of a bit in the whole network decreases as well. Total k-bit that should be transmit is the same in mesh, therefor energy of the NOC decreases, in Eq. (7).

The power equations are expressed as follows [1, 33]:

$$P_{dynamic} = \alpha C V_{DD}^2 f \quad (8)$$

$$P_{total} = P_{dynamic} + P_{static} \quad (9)$$

Where P_{static} and $P_{dynamic}$ are static and dynamic power consumptions; respectively. In the dynamic power consumption, α, C, V_{DD}, and f are activity of chip circuits, capacitance, chip voltage, and bit frequency; respectively. There is also a direct relationship between time and power and energy consumption also increases [34].

In the proposed architecture, Frequency, activity factor, main reference voltage and capacitor's siz is same as mesh architecture. Thus the total power consumption for a single bit in a single link is same as mesh. For h hopcount in a network, with decrease in h parameter, total power consumption decreases, Eqs. (8) and (9).

The energy of one bit in a NoC is calculated through Eq. (6), wherein *h* is the number of hop counts that a bit should transmit to the destination node. In this proposed architecture, a bit transmits fewer hop counts to the destination node; therefore, it has a smaller *h* compared with other architectures. According to Eq. (6), a bit consumes less energy at fewer hop counts (*h*).

Comparison of the total energy consumption in different architectures is shown in Fig. 9, where the total energy consumption augments as NoC expands. This incremental trend is for the reason that there are more hop counts in large NoCs than small ones. In small networks, the amount of energy consumption is not big and the simulation results (Fig. 9) reveal a slight difference in network energy consumption. When NoC grows to about a hundred in the $5 \times 4 \times 4$ network dimension, differences appear and those in energy consumption also increase any time the network

expands. This is because, in bigger networks, differences in diameter are high in relation to a small network. As observed in Eqs. (6) and (7), this difference is much more in a large network with higher than about a hundred nodes compared with small ones. Red lines in diagrams indicate the proposed architecture, presenting less energy consumption than that even from the 3D DB_EP architecture.

In a direct relation, with increase in network size, energy consumption of the network increases, Fig. 9. In each network dimension, the proposed architecture shows an improvement in energy consumption than its counterparts, Fig. 9.

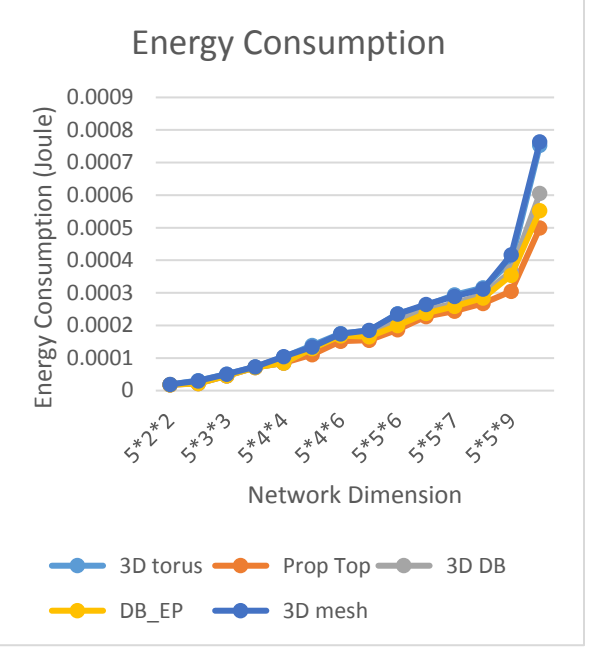
Table 3. Average Node Energy Consumption (J).

Architecture	*Injection Rate 0.05*
3D Mesh	*0.000498318*
3D Torus	*0.00049999*
3D DBG	*0.000438*
3D DB_EP	*0.000386263*
SCMesh	*0.000736218*
Proposed (3 nodes in cluster head)	*0.000274251*

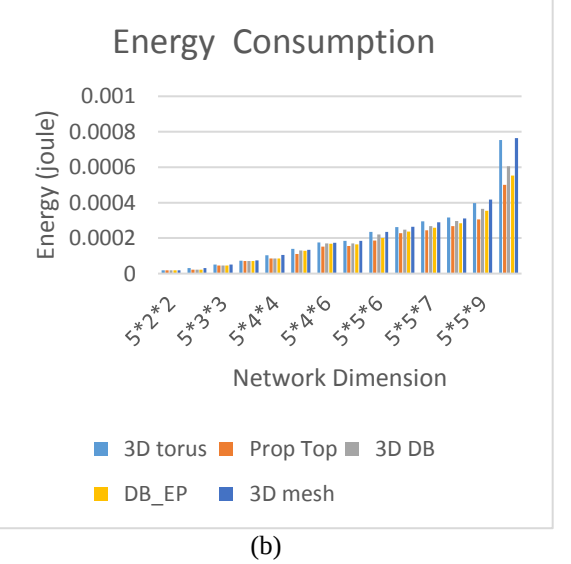
Table 3 outlines average network energy consumption in 0.05 PIR. Following a rise in network PIR of the NoC, energy consumption increases as well. Accordingly, in dimension, each architecture also has a different energy consumption wherein the energy consumption of the proposed architecture is less than that of its counterparts.

The differences of network energy consumption in various network injection rates in $6 \times 6 \times 10$ dimension of different architectures are shown in Fig. 10. When the network injection rate increases, the energy consumption is augmented as well. The line chart and the bar chart in Figs. 10-a and 10-b demonstrate that the proposed architecture has a better consumption of energy among other ones and an average of 32% better energy consumption toward its counterparts.

With increase in network injection rate, network energy consumption increases. Network architectures saturate at injection rate 0.05 flit/node/cycle. In each injection rate, the proposed architecture has better energy consumption. The proposed architecture's energy consumption reveals 32% improvement toward its counterparts, Fig. 10.



(a)



(b)

Fig. 9. (a) Line Chart of Energy consumption of different architectures (b) Bar Chart of Energy consumption of different architectures

Consumption in different network sizes

Energy consumption also has a direct relationship with network latency. Fig.11 shows that network energy consumption increases as network latency is boosted.

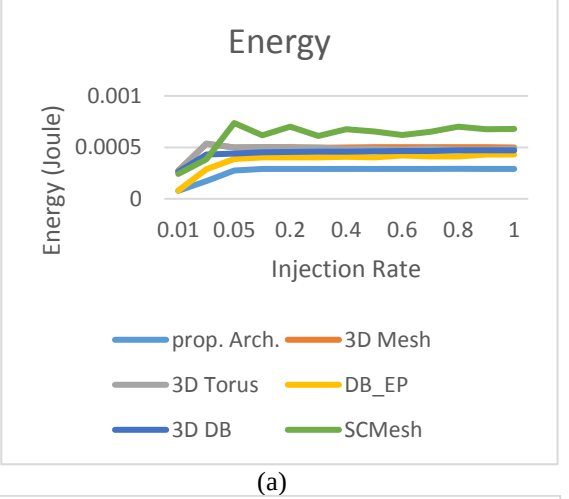
6.3. Summa Cost

The overall cost of the SUMMA in NoC with size $(r \times c) \times h$ is calculated as follows. Assume *x* and *y* are two entry matrices with the size of $m \times n$ having a

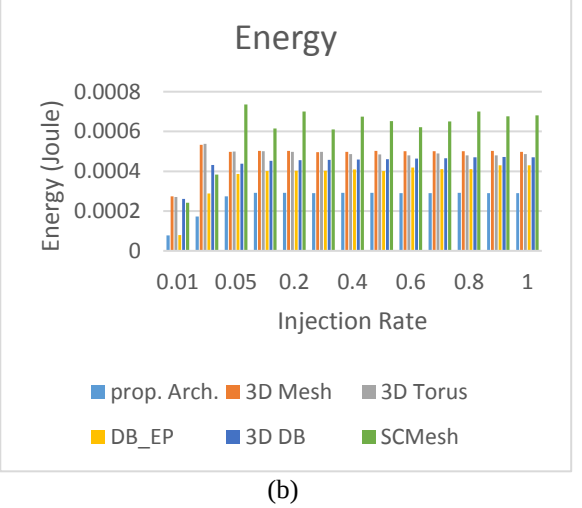
sample network node of $P(I, j, k)$ which is calculated as follows [31]:

$$Cost = \left[2\frac{n_a^3}{P}\right]\gamma + \left[\frac{n}{hb}((\log_2 P) - (\log_2 h)) + 2\log_2 h\right]\alpha + \left[2\left(\frac{\sqrt{P}}{\sqrt{h}} - 1\right) + 3h - 2 + \frac{\gamma}{\beta}h\right]\frac{n_a^2}{P}\beta \quad (10)$$

Wherein, α, β, and γ are communication latency, bandwidth, and computation overload; respectively. b also shows block size of the matrix in the processor. n_a^3 is $m \times n \times$h and na^2 is $m \times n$.



(a)



(b)

Fig. 10. (a) Line Chart of Energy Consumption of different architectures in different injection rates (b) Bar Chart of Energy Consumption of different architectures in different injection rates.

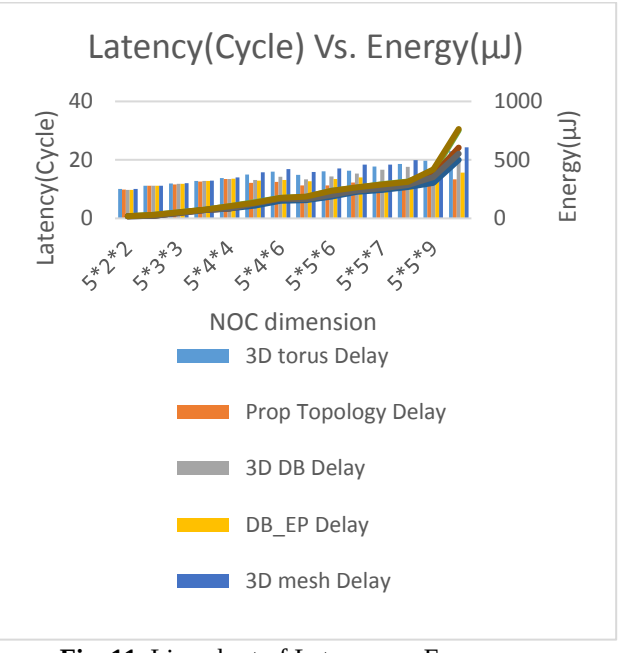


Fig. 11. Line chart of Latency vs. Energy.

The cost of SUMMA for multiplying two 15-by-15 matrices with the number 225 for each one in the proposed NoC architecture is calculated in Table 4, in which the number of processors (P) in a $5 \times 5 \times 5$ network dimension is 125 and the number of block sizes is 3. α represents communication metric and is subsequently calculated according to Noxim NoC simulations in the proposed architecture, suggesting a direct relationship with SUMMA cost over its counterparts. Table 4 shows the cost of SUMMA algorithm in $5 \times 5 \times 5$ architecture.

Table 4. Summa Cost.

Architecture	Costs	α
3D mesh	$27\gamma + 37.84\beta + 225.4984$	24.2994
3D Torus	$27\gamma + 37.84\beta + 213.2182$	22.9761
3D DBG	$27\gamma + 37.84\beta + 179.2896$	19.32
3D DB_EP	$27\gamma + 37.84\beta + 145.7099$	15.7015
Proposed 3D Topology	$27\gamma + 37.84\beta + 123.6096$	13.32

Table 4 outlines the SUMMA cost implemented in $5 \times 5 \times 5$ size in NoC. Moreover, Eq. (10) presents the SUMMA cost in three parts of communication latency, bandwidth, and computation overload. All NoC architectures in Table 4 have the same bandwidth. There is also a difference in computation and communication. In this study, a new NoC architecture is proposed which has a better latency than its counterparts. As well, α

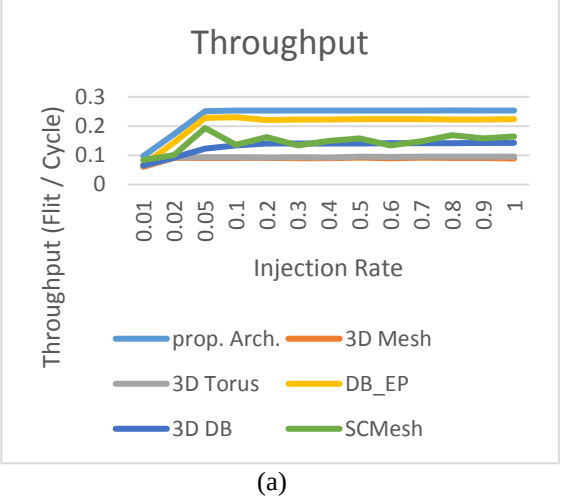
parameter is calculated in the simulator and then presented in the cost equation.

6.4. Throughput

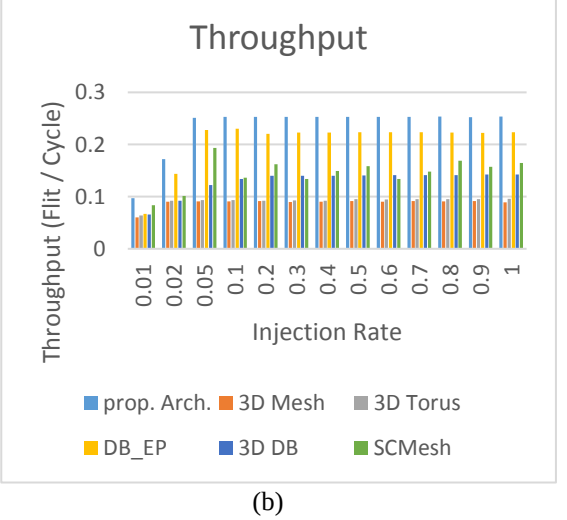
Throughput refers to the number of packets that successfully reach to the destination. Ideal throughput is thus calculated with the following equation [35]:

$$\theta_{ideal} \le \frac{2bB_c}{N} \tag{11}$$

Where, b is bandwidth, N refers to the total number of cores, and B_c represents bisection of channel. In the proposed architecture, node count and network bandwidth is the same as mesh architecture. Bisection channel is the same as mesh architecture and less than its counterparts, Table 6.



(a)



(b)

Fig. 12. (a) Line Chart of throughput of different architectures in different injection rates (b) Bar Chart of throughput of different architectures in different injection rates.

The network throughput in different injection rates is shown in Fig. 12. At an average injection rate (PIR) of 0.05 in network with $6 \times 6 \times 10$, this architecture throughput is going to be in a saturated mode. This proposed architecture also has a better throughput which reaches 13.6% and is better than that of the DB_EP architecture in [11].

With increase in network injection rate, throughput of the network increases. Network architectures saturate at injection rate 0.05 flit/node/cycle. In each injection rate, the proposed architecture has better throughput than its counterparts. The proposed architecture's throughput shows 13.6% improvement toward its counterparts.

Table 5. Average Throughput (Flit / Cycle).

Architecture	Injection Rate 0.05
3D Mesh	*0.0907211*
3D Torus	*0.0930468*
3D DBG	*0.122478*
3D DB_EP	*0.227517*
SCMesh	*0.193334*
Proposed (3 nodes in cluster head)	*0.25093*

Table 5 outlines the average throughput of different NoC architectures at an injection rate of 0.05 and a dimension of $6 \times 6 \times 10$ with 360 cores. Accordingly, it is obvious that the proposed NoC architecture has a better and higher throughput than its counterparts in the same situation.

Network diameter equations are also presented and compared in Table 6 with an example in Table 7. The proposed 3D NoC architecture has a better diameter than its counterparts, indicating that one dimension in this network diameter can be omitted to have a better diameter.

In the DB_EP architecture, a small diameter with long connections leads to much latency and higher energy consumption in big networks. Additionally, the diameter and the connection lengths are important in NoCs. In this topology, with the same short connection length, the diameter is improved and reveals a better performance in big networks than its counterparts.

A comparison of the diameter and the number of connections between different architectures is presented in Table 6, wherein it can be observed that the number of connections is the same as that of mesh and the De-Bruijn topology. If $x = y = z = r$, then $\frac{3N}{2} - r^2 + 2r - 3$ connection yields the same connection numbers in the mesh and the De-Bruijn and less connections in the torus. In this architecture, only some connections in the X dimension are omitted and added to other Z (i.e. the third) dimension, showing a better

diameter which would result in having a better latency and making route transmission from source to destination faster.

Figs. 13 and 14 show the relationship between processor count towards NoC connection count and network diameters in different architectures. With a rise in network processors, the number of connections and diameters increase while the proposed architecture has a better connection and diameter towards its counterparts.

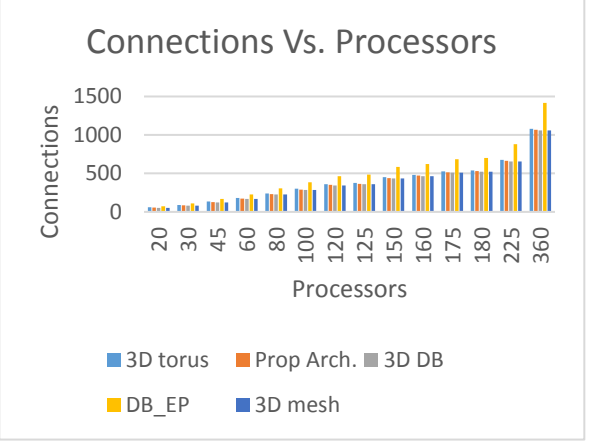


Fig. 13. Line chart of processor count in 3D NOC architectures vs. their connection count.

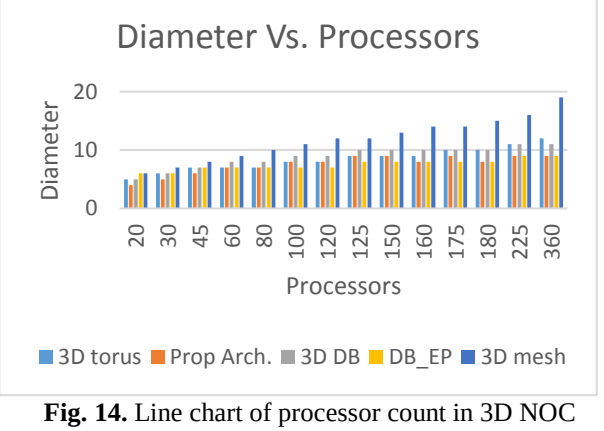


Fig. 14. Line chart of processor count in 3D NOC architectures vs. their connection diameter.

In terms of growth in processors, the diameter is somehow small or equal to this architecture following a significant decrease in the number of connections in the proposed architecture towards the DB_EP architecture, which leads to the small area of the architecture with a better performance than other architectures and lower SUMMA cost. A comparison of NoC architectures through this proposed 3D architecture with 160 and 1000 nodes in the network is presented in Table 7. The network diameter, number of connections, size, and node degree of different 3D architectures are also compared.

7. CONCLUSION AND FUTURE WORKS

A new 3D mesh-based NoC architecture is proposed by applying the star node and De-Bruijn graph. The SUMMA is also implemented on the proposed architecture. This architecture is subsequently compared with 3D mesh, torus, De-Bruijn, DB_EP and SCMesh architectures. It is proved that the given architecture has a small diameter than other ones. Small diameter also leads to less energy consumption and network latency. Simulation results, here, in the Noxim simulator, indicate that the mentioned architecture has a better energy consumption and network latency than its counterparts. An 8.6% improvement in network latency, average 32% improvement in network energy consumption, and 13.6% enhancement in network throughput make this architecture outstanding as a whole.

The matrix multiplication algorithm also faces an improvement in communication time through this architecture. It can be deduced that the proposed architecture outperforms its counterparts with better energy consumption, throughput, and latency.

This architecture can be expanded and designed for a multi-core NoC in future works in a way that its fault tolerance is mentioned. A proper mapping technique in the proposed architecture can be correspondingly implemented to improve the average temperature of each PE.

Table 6. Network Diameter Comparison.

NO	Source	Net. Topology	Dimension	Net. Degree	Net. Diameter	Connections	Bisection	Description
1	[36]	Mesh	3D	4-6	$(r_1-1)+(r_2-1)+(r_3-1)$	$3N-(r_1+r_2+r_3)$	$\sqrt[3]{N}$	$N=r_1\times r_2\times r_3$
2	[36]	Torus	3D	6	$[\frac{r_1}{2}]+[\frac{r_2}{2}]+[\frac{r_3}{2}]$	3N	$2\sqrt[3]{N}$	$N=r_1\times r_2\times r_3$
3	[10]	3D DB	3D	6	$\log r_1+\log r_2+(r_3-1)$	$3N-(r_1+r_2+r_3)$	$\sqrt[3]{N}$	$N=r_1\times r_2\times r_3$
4	[11]	DB_EP	3D	6	$\log r_1+\log r_2+2$	3N-(r1+r2+r3)+N	$\sqrt[3]{N}+1$	$N=r_1\times r_2\times r_3+r_3$
6	[13]	SCMesh	3D	3-8	5+(r-1)	3N-3r+18	3r+12	$N=6\times 6\times r$
7	Proposed	Proposed	3D	6	$[\log Y+Z+1]$	$3N-(X+Y)$	$\sqrt[3]{N}$	$N=X\times Y\times Z$

TABLE 7. Network Diameter for $6 \times 6 \times 4$ of 144 nodes network.

NO	Source	Net. Topology	Dimension	Net. Degree	Processing Element	Net. Diameter	Connections	Bisection	Description
1	[36]	Mesh	$6 \times 6 \times 4$	4-6	144	13	416	5	$N = 144$
2	[36]	Torus	$6 \times 6 \times 4$	6	144	8	432	10	$N = 144$
3	[10]	3D DB	$6 \times 6 \times 4$	6	144	6	416	5	$N = 144$
4	[11]	DB_EP	$6 \times 6 \times 4$	6	144	7	560	6	$N = 144$
6	[13]	SCMesh	$6 \times 6 \times 4$	3-8	144	10	434	24	$N = 144$
7	Proposed	Proposed	$6 \times 6 \times 4$	6	144	6	420	5	$N = 144$

REFERENCES

[1] Feero, B. Stanley, & P. P. Pande, "**Networks-on-chip in a three-dimensional environment: A performance evaluation**," IEEE *Transactions on Computers*, 2009.

[2] ABBAS, Assad, et al., "**A survey on energy-efficient methodologies and architectures of network-on-chip**," *Computers & Electrical Engineering*, 2014.

[3] MANNA, Kanchan, et al., "**Thermal-aware application mapping strategy for network-on-chip based system design**," *IEEE Transactions on Computers*, 2018.

[4] S. Azampanah, A. Khademzadeh, N. Bagherzadeh, M. Janidarmian, & R. Shojaee, "**Contention-aware selection strategy for application-specific network-on-chip**," *IET Computers & Digital Techniques*, 2013.

[5] S. Forghani, N. Habibi, & M. Firoozbakht, "**Network security metric based on attack duration**," *In 2015 2nd IEEE International Conference on Knowledge-Based Engineering and Innovation (KBEI),* 2015.

[6] A. More, V. Pano, & B. Taskin, "**Vertical Arbitration-Free 3-D NoCs**," IEEE *Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 2018.

[7] R. K. Saini, & M. Ahmed, "**2D hexagonal mesh Vs 3D mesh network on chip: A performance evaluation**," *International Journal of Computing and Digital Systems*, 2015.

[8] M. D. Schatz, R. A. Van de Geijn, & J. Poulson, "**Parallel matrix multiplication: A systematic journey**," *SIAM Journal on Scientific Computing*, 2016.

[9] R. Dash, A. Majumdar, V. Pangracious, A. K. Turuk, & J. L. Risco-Martín, "**ATAR: An Adaptive Thermal-Aware Routing Algorithm for 3-D Network-on-Chip Systems**," *IEEE Transactions on Components, Packaging and Manufacturing Technology,* 2018.

[10] R. Sabbaghi-Nadooshan, M. Modarressi, & H. Sarbazi-Azad, "**A Novel De Bruijn Based Mesh Topology for Networks-on-Chip**," *In VLSI. IntechOpen, BoD–Books on Demand*, 2010.

[11] Y. Chen, J. Hu, X. Ling, & T. Huang, "**A novel 3D NoC architecture based on De Bruijn graph**," *Computers & Electrical Engineering*, 2012.

[12] W. Gao, & P. Zhou, "**Customized high performance and energy efficient communication networks for AI chips**," *IEEE Access*, 2019.

[13] R. Poovendran, & S. Sumathi, "**An area-efficient low-power SCM topology for high performance network-on Chip (NoC) architecture using an optimized routing design**," *Concurrency and computation: practice and experience*, 2019.

[14] N. Viswanathan, K. Paramasivam, & K. omasundaram, "**Exploring Hierarchical, Cluster based 3D Topologies for 3D NoC**," *Procedia Engineering*, 2012.

[15] A. Touzene, "**On all-to-all broadcast in dense Gaussian network on-chip**," *IEEE Transactions on Parallel and Distributed Systems*, 2015.

[16] I. Pires, M. Alves, & L. Albini, "**Expansible Network-on-Chip Architecture**," *Advances in Electrical and Computer Engineering*, 2018.

[17] B. Debajit; J., Niraj K. "**Analytical modeling of the SMART NoC,**" *IEEE Transactions on Multi-Scale Computing Systems*, 2017.

[18] K., Tushar, et al. **"Breaking the on-chip latency barrier using SMART,"** *IEEE, 19th International Symposium on High Performance Computer Architecture (HPCA)*. IEEE, 2013.

[19] P. Yang, & Q. Wang, "**Heterogeneous honeycomb-like NoC topology and routing based on communication division**," *International Journal of Future Generation Communication and Networking*, 2015.

[20] P. Lotfi-Kamran, A. M. Rahmani, M. Daneshtalab, A. Afzali-Kusha, & Z. Navabi, "**EDXY–A low cost congestion-aware routing algorithm for network-on-chips**," *Journal of Systems Architecture*, 2010.

[21] N. Viswanathan, K. Paramasivam, & K. Somasundaram, "**Performance and Cost Metrics Analysis of a 3D NoC Topology using Network Calculus**," *Applied Mathematical Sciences*, 2013.

[22] A. Ahmadinia, & A. Shahrabi, "**A highly adaptive and efficient router architecture for network-on-chip**," *The Computer Journal*, 2011.

[23] D. Demirbas, I. Akturk, O. Ozturk, & U. Güdükbay, "**Application-specific heterogeneous network-on-chip design**," *The Computer Journal*, 2014.

[24] S. E. Bae, T. W. Shinn, & T. Takaoka, "**A faster parallel algorithm for matrix multiplication on a mesh array**," *Procedia Computer Science*, 2014.

[25] M. Hosseinabady, M. R. Kakoee, J. Mathew, & D. K. Pradhan, "**Low latency and energy efficient scalable architecture for massive NoCs using generalized de Bruijn graph**," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, 2011.

[26] O. Collins, F. Pollara, S. Dolinar, & J. Statman, "**Wiring Viterbi decoders (splitting de Bruijn graphs)**," *The Telecommunications and Data Acquisition Progress Report*, 1989.

[27] A. Louri, H. & Sung, "**An efficient 3D optical implementation of binary de Bruijn networks with applications to massively parallel computing**," *In IEEE Proceedings of Second International Workshop on Massively Parallel Processing Using Optical Interconnections*, 1995.

[28] M. R. Samatham, & D. K. Pradhan, "**The de Bruijn multiprocessor network: a versatile parallel processing and sorting network for VLSI**," *IEEE Transactions on Computers*, 1989.

[29] P. Faizian, M. A. Mollah, X. Yuan, Z. Alzaid, S. Pakin, & M. Lang, "**Random Regular Graph and Generalized De Bruijn Graph with k-Shortest Path Routing**," *IEEE Transactions on Parallel and Distributed Systems*, 2018.

[30] D. Hoxha, "**Sparse Matrices and Summa Matrix Multiplication Algorithm in STAPL Matrix Framework**," *Doctoral dissertation*, 2016.

[31] M. Schatz, J. Poulson, & R. van de Geijn, "**Parallel Matrix Multiplication: 2D and 3D**", *FLAME Working Note No. 62, The University of Texas at Austin*, 2012.

[32] Y. S. Yang, H. Deshpande, G. Choi, & P. V. Gratz, "**SDPR: Improving Latency and Bandwidth in On-Chip Interconnect Through Simultaneous Dual-Path Routing**," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 2018.

[33] T. Maqsood, S. Ali, S. U. Malik, & S. Madani, "**A.Dynamic task mapping for network-on-chip based systems**," *Journal of Systems Architecture*, 2015.

[34] N. Jafarzadeh, M. Palesi, A. Khademzadeh, & A. Afzali-Kusha, "**Data encoding techniques for reducing energy consumption in network-on-chip**," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, 2014.

[35] H. Matsutani, M. Koibuchi, & H. Amano, "**Tightly-coupled multi-layer topologies for 3-D NOCs**," *In International Conference on Parallel Processing (ICPP), IEEE*, 2007.

[36] K. Hwang, & N. Jotwani, "**Advanced Computer Architecture**," 3e. *McGraw-Hill Education*, 2016.