Automatika

Journal for Control, Measurement, Electronics, Computing and Communications

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/taut20

# Dynamic low power management technique for decision directed inter-layer communication in three dimensional wireless network on chip

T. R. Dinesh Kumar & A. Karthikeyan

**To cite this article:** T. R. Dinesh Kumar & A. Karthikeyan (2023) Dynamic low power management technique for decision directed inter-layer communication in three dimensional wireless network on chip, Automatika, 64:4, 1280-1295, DOI: 10.1080/00051144.2023.2261088

**To link to this article:** https://doi.org/10.1080/00051144.2023.2261088

Published online: 06 Oct 2023.

AUTOMATIKA
2023, VOL. 64, NO. 4, 1280–1295
https://doi.org/10.1080/00051144.2023.2261088

Taylor & Francis
Taylor & Francis Group

OPEN ACCESS  Check for updates

# Dynamic low power management technique for decision directed inter-layer communication in three dimensional wireless network on chip

T. R. Dinesh Kumar[a] and A. Karthikeyan[b]

[a]Department of Electronics and Communication Engineering, Vel Tech High Tech Dr.Rangarajan Dr.Sakunthala Engineering College, Avadi, India; [b]Department of Electronics and Communication Engineering, Vel Tech Multi Tech Dr.Rangarajan Dr.Sakunthala Engineering College, Avadi, India

**ABSTRACT**

3D ICs, a novel technology, might significantly impact multicore NoCs with hundreds or thousands of processing components on a single chip. Multiple 2D chips can be stacked vertically to create multiple active processing elements at various levels. Adding active device layers to 3D ICs can enhance system performance, increase functionality, and increase packing density. New architectural and IC technology advancements hinder energy-efficient design research. Achieving a balance between chip power and performance is crucial. This paper describes the “Dynamic Low Power Management Method in 3DWiNoC” (DLPM 3DWiNoC) architecture, which enables self-organized, centrally managed service management using Smart Master Agents. The approach utilizes SMA’s ODA DD module for self-organized, centrally managed service management. To improve power regulation, data flow across vertical interconnects (TSVs) is reconfigured based on a dynamic evaluation of channel link use. SMA aims to reduce congestion by increasing connection utilization through high-frequency, bi-directional vertical channels via TSVs. The suggested system is modeled in MATLAB Simulink. Compared to 3D stacking, TSV stacking of vertical interconnects with the SMA method ensures low parasitic (latency and power) and higher bandwidth with higher vertical wire densities. Experimental results show that the proposed architecture decreases area overhead by 5%-7%, network latency by 12%-15%, and NoC power consumption by 15%-20% compared to the present multi-NoC design.

**ARTICLE HISTORY**
Received 8 June 2023
Accepted 16 September 2023

**KEYWORDS**
Decision Directed Communication; centralized service management; vertical channels; inter-layer communication; optimize power control; link utilization

## Introduction

By cramming several cores onto a single chip, the network-on-chip (NoC) [1] paradigm has evolved as a new technology. The physical manifestation of communication channels forced by metal wires with long interconnects [2] presents a performance barrier both in terms of power and latency as the number of cores in a single device keeps growing. Moreover, frequent packet transmission puts a tremendous amount of strain on the constrained floor layout in typical 2D integrated circuits (ICs), which lowers performance in NoC architectures. The International Technological Roadmap for Semiconductors (ITRS) states that in order to provide an extraordinary performance boost, new interconnect models are required. An emerging technology known as three-dimensional integration circuits (3D ICs) [3] allows for the vertical stacking of several 2D chip layers using layer-to-layer interconnections that comprise multiple layers of active processing elements. Many active device layers in 3D ICs have the potential to improve system performance and can lead to higher functionality and packing densities. A few benefits of 3D NoC include shorter connecting wires, a smaller network’s diameter, enhanced memory bandwidth, and improved form factors that result in higher packing densities and a smaller footprint. A significant issue arises when creating an interconnection infrastructure for multi-core, multi-layer systems. A “Dynamic Low Power Management Method for 3D WiNoC Architecture (DLPM 3DWiNoC)” is what this paper’s proposed effort intends to create. The model makes it possible to combine two new paradigms, specifically WiNoCs in a 3D IC environment. The objective is to construct a highly effective multi-core, multi-layer WiNoC to maximize the design and fabrication of on-chip miniature antennas operating in the gigahertz range with high-density vertical NoC links through silicon vias (TSVs), to dynamically enable effective low-power intra- and inter-chip communication [4]. The model takes advantage of the benefits of using cutting-edge mm-wave wireless links [5] that operate in the 60 GHz band in multichip systems to reduce the downsides of multi-hop communication. The DLPM 3DWiNoC model is characterized by its capacity to create and construct a reconfigurable 3D WiNoC. The “Smart Master Agent”, a centrally managed wireless router, has the ability to

**CONTACT** T. R. Dinesh Kumar ✉ trdineshkumar@velhightech.com Assistant Professor, Department of Electronics and Communication Engineering, Vel Tech High Tech Dr Rangarajan Dr Sakunthala Engineering College, Avadi, Chennai, Tamilnadu, 600062, India

dynamically change the vertical channel bandwidth in accordance with application usage patterns. The goal is to improve communication inside and between layers. (ii) Through Silicon Vias (TSV) buses that connect adjacent stacked wafers in order to take advantage of vertical linkages that run at higher frequencies and have the capacity to dynamically change direction [6] based on the real-time bandwidth needed for efficient inter-layer communication. MATLAB Simulink is used to model the suggested system. By gauging performance using benchmark metrics for 3D NoC structures, a real traffic pattern in a cycle-accurate simulation is implemented. According to comparative research, the suggested TSV stacking of vertical interconnects with SMA method ensures low parasitic (i.e. low latency and power) with high densities of vertical wires allowing high bandwidth than its existing counterparts. The separation of the functional units supporting heterogeneous capacity into distinct levels is a common result of the 3D stacking [7] of numerous chips. Also, the suggested approach considerably reduces the inter-layer wire length, making 3D stacking appealing for future high-performance computer systems. Different levels are connected using vertical TSVs. However, the high-density vertical TSVs coupled with the SMA method make the entire system entirely compatible with precise post-layout timing, area, and power analyses to effectively connect layers running at various clock frequencies, making 3D WiNoCs a very promising method. The remaining portions of the text are arranged as follows: The related work is presented in Section II, followed by a detailed explanation of the DLPM 3DWiNoC system structure in Section III, the simulation results and analysis in Section IV, and the conclusion of the study and a discussion of future directions in Section V.

## Related works

All across the world, 3D NoC research is actively being conducted. Wafer-level stacking, die-on-wafer stacking, transistor stacking, and chip stacking are some of the well-known 3D NoC research topics [8]. With the reduction in feature size, 3D integrated circuits have established themselves as an appealing solution for overcoming the limitations of interconnect scaling and providing a chance to maintain CMOS technology's performance advancements. In this part, various recent research projects are discussed. These projects mostly use 2D and 3D NoC architecture, supported by on-chip wireless communication. In recent NoC research literature, many 3D-NoC structures have been proposed. Each of these architectural designs has advantages and disadvantages of its own. A few 2D and 3D-NoC designs are briefly illustrated in the section that follows. A well-known traditional 2D mesh-based NOC design is discussed by the author in [9]. Short interswitch cables and a regular structure are two characteristics of this architecture. It is made out of an MXN mesh network of switches that connect IP blocks. However, the variety of network structures is now restricted due to the growing number of cores per chip in 2D mesh. It is susceptible to communication bottlenecks and energy inefficiency because of the additional hops due to wire-length restrictions and layout issues. By comparing each network's zero-load latency and power use, the work in [10] contrasts 2D MESH architectures with their 3D counterparts. Although its evaluation in 3D NoCs is favourable, it does not measure other important performance metrics utilizing real-time traffic patterns. The study in [11] highlights the difficulties in manufacturing 3D ICs and different fabrication techniques for optimizing functionality, power, density, and noise reduction in 3D ICs. By lowering the network width and communication distance during transmission, the 3D mesh architecture greatly outperforms the 2D mesh architecture in terms of communication performance while consuming less power. In [12], the author explains a well-known stacked 3DMesh architecture in which the cores are organized into many layers and piled on top of one another. Negligible inter-wafer distance in 3D chips and a longer crossbar in each switch due to an increase in the number of ports per switch are disadvantages of such a design. Negligible inter-wafer distance in 3D chips and a longer crossbar in each switch due to an increase in the number of ports per switch are disadvantages of such a design. Each switch in a ciliated 3D mesh architecture [13] has a maximum of $5 + k$ ports, including one for each cardinal direction, two for up and down, and one for each of the k IP blocks. This architecture offers less overall bandwidth than a symmetric 3D mesh but is typically more energy efficient due to the many IP cores per switch and decreased connection. Several other studies also take into account mapping tree-based interconnection networks [14] established in 2D NoC onto a multilayer 3D SoC to streamline wire routing and shorten the longest inter-switch cable length to improve functionality and boost system performance. Nonetheless, this may result in less energy loss and smaller overheads for the area. Despite all of the benefits of 3D-NoC, this design paradigm has quite a few drawbacks, including difficulty with electrical modelling, power management, thermal management, and fabrication problems such as coupling between TSVs [15,16]. Despite the significant benefits of 3D integration, significant obstacles, including thermal mitigation and efficient interconnect power management problems, still exist. This work, DLPM 3DWiNoC, presents an alternative solution that combines the advantages of two new areas, such as 3D wireless NoC integrated with TSVs, to address the issues of excessive latency and energy dissipation in a NoC. In the section below, the proposed work is illustrated in detail.

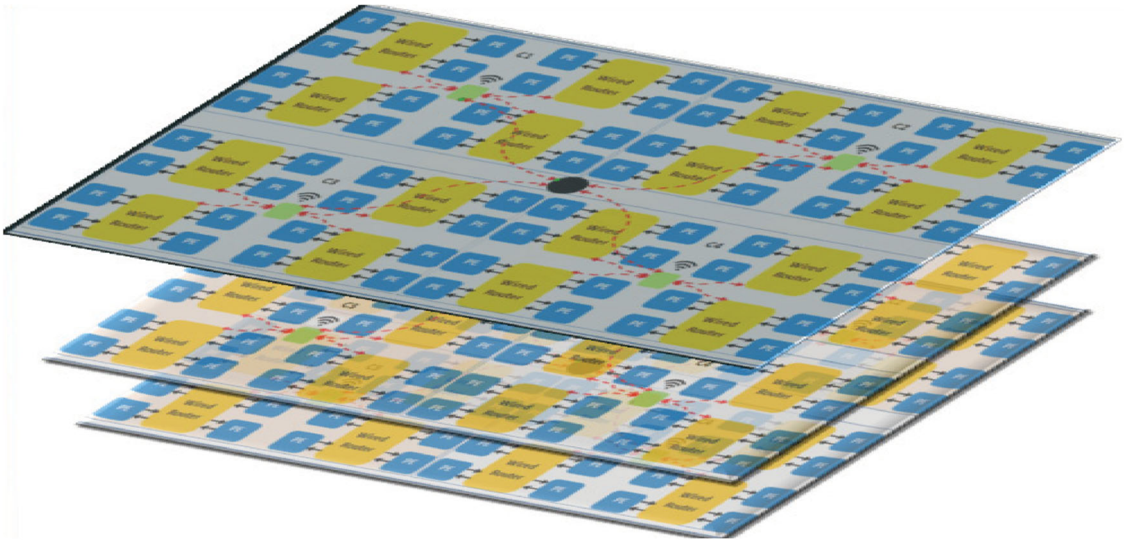

**Figure 1.** The DLPM 3DWiNoC system model.

## Dynamic low power management technique for decision directed inter-layer communication in three dimensional wireless network on chip system framework

The DLPM 3DWiNoC system under consideration offers design in a vertical dimension, giving designers a great deal of flexibility in selecting an on-chip network topology. As seen in Figure 1, this architecture comprises 8 × 8x8 multi-layer, multi-core IP blocks connecting wired and wireless routers. The system integrates eight layers, linking them with TSVs that span the chip's entire vertical dimension. The suggested model divides each layer into four clusters, each of which has 16 cores. Every cluster in the system has a wireless router (known as an "intelligent head agent") that connects four wired routers through a wireless link and four wired routers with four cores each. The Smart Master Agent, a wireless router that connects all 4 IHAs (wireless routers) in each cluster through a wireless link, is located at the centre of the layer. Overall, the suggested approach integrates functional elements, including cores and wired and wireless routers, onto a single chip using 3D stacking, as shown in Figure 1. High-performance and high-bandwidth intra – and inter-layer communication are more appealing to 3D stacking thanks to the potential of heterogeneous integration.

Vertical through-silicon vias (TSVs) are used to connect the layers, greatly reducing the length of the inter-layer wire. The TSVs are positioned in the chip's centre to make routing easier.

These groupings or buses of vertical wires are called buses. Nine 64-bit vertical TSV links between layers are the maximum number to address chip area and manufacturing cost constraints. The suggested method employs a dynamic power management methodology that improves power characteristics to primarily modify the frequency of vertical channels in accordance with traffic load. Via vertical TSVs, the bi-synchronous FIFOs take advantage of high-speed inter-layer communication. Table 1 lists the nomenclature used in the proposed model.

*Packet Structure:* The packets are divided into smaller pieces known as flits. Figure 2 shows the DLPM 3DWiNoC model's packet structure.

Each packet is made up of three parts: the header, the route, and the content. The packet format, as shown in Figure 2, comprises nine fields, including Source (Sc), Destination (Dc), Operation (Op), Type, Role, Priority, and Size.

**Table 1.** Nomenclature of DLPM 3DWiNoC model.

| Notation | Description |
|---|---|
| PE$^{i}_{id}$ | Processing element identity |
| PE$^{i}_{load}$ | Processing element load |
| PE$_{P\_value}$ | Processing element priority value |
| WR$_{id}$ | Wired router identity |
| WRL$_{status}$ | Wired router load status |
| WRRAflit | Wired router ready to accept flit |
| WRBflit | Wired router busy flit |
| TWRL$_{status}$ | Threshold wired router load |
| SMAflit | Smart Master Agent header flit |
| SMA$_{id}$ | Smart Master Agent identity |
| SMA$_{TS}$ | Smart Master Agent time stamp |
| SMA$_{node\_status}$ | Smart Master Agent node status |
| SMA$_{Link\_util}$ | Smart Master Agent link utilization |
| IHAflit | Intelligent Head Agent header flit |
| IHA$_{id}$ | Intelligent Head Agent identity |
| IHA$_{TS}$ | Intelligent Head Agent time stamp |
| IHA$^{i}_{node\_status}$ | Intelligent Head Agent node status |

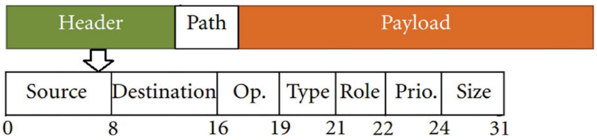

**Figure 2.** Packet Structure of DLPM 3DWiNoC.

*Data Classification and Scheduling Mechanism:* The process assigns the data flow a priority level, either low or high. The priority field in the header flit is set to either 0 or 1 to indicate normal or emergency flow, respectively. Crucial messages are regarded as high-priority emergency data flows that call for high reliability, ample bandwidth, and quick data delivery. While the typical flow necessitates the data's timely delivery to its destination, the suggested approach uses stringent scheduling discipline along with a probabilistic priority (SPP) queuing procedure that allows service differentiation based on the volume of traffic that the queue is able to handle. In order to determine which packet should be forwarded across the output link during a network cycle, the SPP algorithm examines the queues by looking at the bandwidth ratio that has been allotted to each queue. When there are packets ready for transmission, the queues share the available bandwidth. In order to process high and low-priority traffic classes, the wired router divides data flow into separate HP and LP queues. The weights and the proportion of packets waiting in each queue determine how many packets are added to and removed from each queue proportionally based on the priority. Let $h$ and $l$ stand for the proportion of HP and LP packets that are currently in the queues for each, which is determined as follows:

$$h = QL_{\mathrm{HP}}/(QL_{LP} + QL_{\mathrm{HP}}) \tag{1}$$

$$l = QL_{\mathrm{LP}}/(QL_{\mathrm{LP}} + QL_{\mathrm{HP}}) \tag{2}$$

where $\mathrm{QL_{HP}}$ and $\mathrm{QL_{LP}}$ stand for the dynamically-changing number of packets occupied in the HP and LP queues, respectively. The amount of packets from each queue that will be de-queued is determined by their access ratio and is computed as follows:

$$w_{\mathrm{HP}}h : w_{\mathrm{LP}}l \tag{3}$$

where "$w$" stands for a constant that denotes weight. The user-defined weights ($w_{\mathrm{HP}}$ and $w_{\mathrm{LP}}$) given to the HP and LP queues are such that $w_{\max} > w_{\mathrm{HP}} > w_{\mathrm{LP}} > 0$. The maximum weight is represented by $w_{\max}$. As a result, it is guaranteed that HP packets will always take precedence over LP packets for any random data flow.

## Specification of dynamic low power management technique for 3D WiNoC architecture in multi-layer multi-core WiNoC system

The specification and functioning of the various phases of the proposed model are thoroughly explained in this section. The 8 × 8 × 8 WiNoC topology system that has been developed divides the system into clusters to maximize data transfer across cores. Each layer is split into four clusters by it. Each cluster has 16 cores, 4 wired routers, and one wireless router that is accessible from all wired routers because it is positioned in the centre of the cluster. In contrast to alternative topologies, the proposed technique only uses wired interconnects to connect IP cores to wired routers. The wireless link connecting all wired routers to the wireless router installed in the cluster's core. The main focus of this study is on the TSVs and the Smart Master Agent's basic functionality (centralized service management). While the operation of devices like wired routers and Intelligent Head Agent is outside the purview of this work.

## Phase of DLPM_3DWiNoC model

The following phases make up the main functionality of the 3DWiNoC model's Dynamic Low Power Management Technique:

- Phase I: Reliable SMA Selection among Stacked Layers and the Function of SMA
- Phase II: Reconfigurable Vertical Inter-connects Using TSVs for Optimum Power Regulation
- Phase III: Self-organized Centralized Service Management Using SMA's ODA-DD Module

### *Phase I: reliable SMA selection among stacked layers and the function of SMA*

Each layer's central location has a wireless router installed that serves as a master node, also known as the Smart Master Agent (SMA). In order to use intra-layer communication using horizontal wireless links, the SMA is equipped with a mm Wave wireless transceiver and an on-chip zigzag Ansoft HFSS antenna. It also has an Optimised Data Communication for Decision Directed Inter-Layer Communication (ODA DD) module that enables it to provide self-organized centralized service to other layers. Figure 3 shows that the SMA node uses vertical TSVs and the ODA DD module to provide service to other stacked layers. Wireless networks link up all IHAs in each layer.

The Smart Master Agent's (SMA) job description is to:

- Self-organize the network using ODA DD module capabilities for stacked layer power management
- Makes good use of the link and reduces congestion by using bi-directional, bi-synchronous vertical channels that run at high frequency through TSVs.
- Changes the direction of the channel across SMAs in the surrounding layers based on real-time link usage data.

### *The reliable SMA selection among stacked layers involves the following mechanism*

In order to register their current load and status information in their respective routing tables, the member core or Processing Element (PE), WRs, IHAs, and SMAs continuously broadcast the header flit ($\mathrm{PE_{flit}}$,

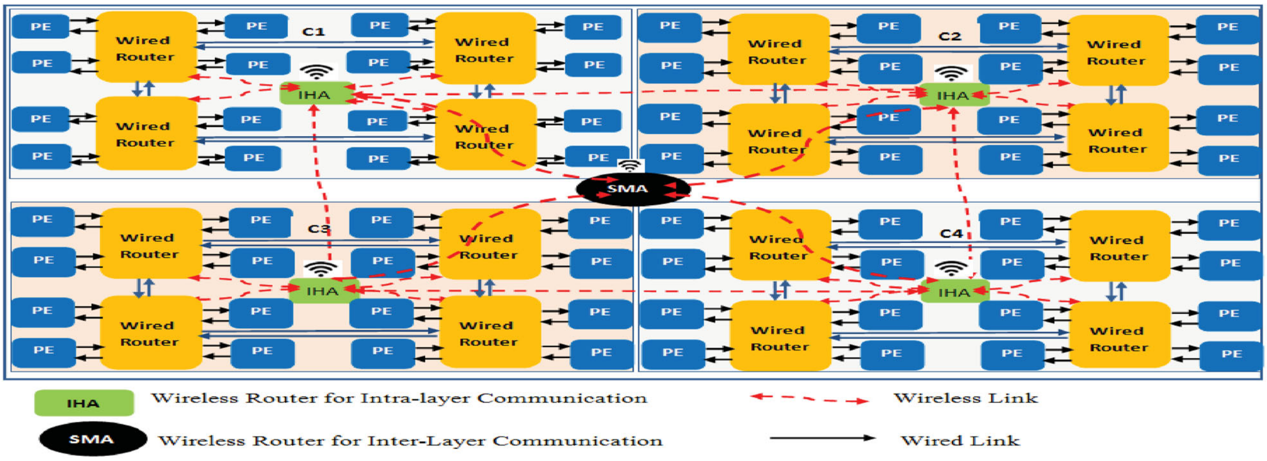
PE
Wired Router
IHA
SMA
C1
C2
C3
C4
IHA Wireless Router for Intra-layer Communication
SMA Wireless Router for Inter-Layer Communication
Wireless Link
Wired Link

**Figure 3.** Functional units of DLPM_3DWiNoC Model.

**Table 2.** Wired routing table.

| $WR_{id}$ | $WRL_{status}$ | $WRL_{TS}$ | $PE^{i}_{id}$ | $PE^{i}_{load}$ | $PE^{i}_{TS}$ | $PE^{i}_{P\_value}$ |
|---|---|---|---|---|---|---|
| $NWR^{i}_{id}$ | $NWRL^{i}_{status}$ | $NWRL^{i}_{TS}$ | $PE^{i}_{id}$ | $PE^{i}_{load}$ | $PE^{i}_{TS}$ | $PE^{i}_{P\_value}$ |

**Table 3.** Header routing table.

| $NIHA_{id}$ | $NIHA_{TS}$ | $MWR^{i}_{id}$ | $MPE^{i}_{id}$ | $IHA^{i}_{node\_status}$ |
|---|---|---|---|---|

$WR_{flit}$, $IHA_{flit}$, and $SMA_{flit}$) to their neighbours in the network (Wired Routing Table, Header Routing Table, Master Routing Table). The following steps are involved in the probe or header flit transmission procedure between network nodes:

Step1: The Wired Router (WR) receives, extracts, and adds an entry to its routing database called the Wired Routing Table after receiving a probe signal from the member PEs on a regular basis (WRT). Each cluster's Wired Router continuously broadcasts the "$WR_{flit}$" header flit to its neighbour WRs, which are connected by wired medium, as well as to the IHA (connected through wireless link). The following details are included in the "$WR_{flit}$":

$$WR_{flit} = [WR_{id}, WRL_{status}, ((PE^{i}_{id}, PE^{i}_{load}, PE_{TS}, PE_{P\_value}), \ldots)]$$

As seen in Table 2, the neighbouring WR's add the entry to its WRT after receiving the $WR_{flit}$. Hence, the PE information of each WR's neighbouring WRs is contained in the WRT of each WR.

These fields are available for entry in WRT: ($PE^{i}_{id}$, $PE^{i}_{load}$, $PE_{TS}$, $PE_{P\_value} \ldots$) $WRT_{fields}$. The WR keeps track of a buffer load value ($WRL_{status}$) that represents the load status at any given time. Also, the $NWR^{i}_{id}$ shows the ID of the neighbouring wired router, the $NWRL^{i}_{status}$ shows the load status of the neighbouring wired router, and the $NWRL^{i}_{TS}$ shows the time stamp of the neighbouring wired router. The $NWR^{i}_{TS}$ list shows the neighbouring wired router's core ids, load, time stamp, and priority for data transmission.

Step 2: All of the WR in the same cluster send their "$WR_{flit}$" to the IHA. It enters the member's data in its header routing table after receiving. The IHA transmits the Smart Master Agent and its IHA neighbours, the "$IHA_{flit}$" header flit (SMA). The following details can be found in the "$IHA_{flit}$" flit: $IHA_{flit}$ is defined as

$$[IHA_{id}, IHA_{TS}, MWR^{i}_{id}, MPE^{i}_{id}, \text{and } IHA^{i}_{node\_Status}]$$

where $IHA_{id}$ denotes the $IHA_{id}$, $IHA_{TS}$ denotes the IHA time stamp, and $IHA^{i}_{node\_Status}$ denotes each of their IHA node statuses. After receiving the $IHA_{flit}$ from the neighbour, that IHA extracts the sender's information and adds the entry to its header routing table (HRT). Each IHA has a HRT table, which is shown in Table 3, and comprises a list of neighbouring IHAs, information on each IHA's member wired router, and node status information. These fields are available for entry in HRT:

$$HRT_{fields} = [IHA^{i}_{node\_status}, NIHA_{id}, NIHA_{TS}, MWR^{i}_{id}, MPE_{id}]$$

where $NIHA_{id}$ stands for "neighbour $IHA_{ID}$", and $NIHA_{TS}$ for "neighbour IHA time stamp", "neighbour IHA member $WR_{ID}$", "neighbour IHA member $PE_{ID}$", and "neighbour IHA node status information".

Similar to this, all IHAs communicate with one another via the "$IHA_{flit}$" transmission to transmit information about their members and node status, which is then added to and kept in each HRT.

Step 3: Each IHA transmits its $IHA_{flit}$ to a central Smart Master Agent (SMA), which registers it in the Master Routing Table (MRT). To register the list of IHAs with their member wired routers' load status data, the SMAs in each layer communicate $SMA_{flit}$ to SMAs in other layers. The following details are found in the

"$\mathrm{SMA}_{\mathrm{flit}}$" flit:

$$\begin{aligned}\mathrm{SMA}_{\mathrm{flit}} = {} & [\mathrm{SMA}_{\mathrm{id}},\mathrm{SMA}_{\mathrm{TS}},\mathrm{SMA}_{\mathrm{node\_status}},\\ & \mathrm{SMA}_{\mathrm{Link\_util}},]\\ & [(\mathrm{MIHA}^{i}_{\mathrm{id}}, \mathrm{MIHA}^{i}_{\mathrm{node\_status}},\\ & \mathrm{MWR}^{i}_{\mathrm{id}}, \mathrm{MPE}^{i}_{\mathrm{id}}, ) \ldots],\end{aligned}$$

where $\mathrm{SMA}_{\mathrm{id}}$ is the $\mathrm{SMA}_{\mathrm{id}}$, $\mathrm{SMA}_{\mathrm{node\_status}}$ and $\mathrm{SMA}_{\mathrm{Link\_util}}$ are the node status and link usage of SMA, along with its member data. Similar to this, the SMA in each layer transmits information about itself and its members via "$\mathrm{SMA}_{\mathrm{flit}}$", which is then added to and kept in each corresponding MRT as shown in Table 4.

The SMA in each layer can now choose the best SMA and IHA to reach the destination core during data transmission by using the information in MRT. Each node has the capacity to choose the neighbour nodes that are the most trustworthy using the information in the routing table. This guarantees that each node will only submit service queries to reliable neighbours (RNs), who may then pass them along to their RNs.

The proposed model says that there are three situations in which a packet must be sent from a source core to a destination core:

Case 1: Source core and destination core exists in same subnet.

Depending on the traffic load patterns, data is sent from the source to the destination using either wired or wireless links. In this case, the source core is in one subnet and the destination core is in the same subnet on the same layer.

Case 2: Source core and destination core exists in different subnet but in same layer.

IHA is used for wireless intra-layer communication between a source and a destination when the source core is in one subnet and the destination core is in a different subnet within the same layer.

Case 3: Source core and destination core exists in different subnet in different layer.

IHA is used for wireless intra-layer communication, and SMA is used for inter-layer communication when data transmission between source and destination (i.e. in this scenario, the source core resides in one subnet while the destination core exists in another subnet in a different layer) is required. The path information is stored in the path flit "pflit". The Destination Core DC starts the resource reservation operation along the reverse path after receiving the "pflit". The source core Sc is now admitting data traffic along the chosen route to the destination.

### *Phase II: re-configurable vertical inter-connects using TSVs for optimized power control*

Each layer in the suggested model has a single SMA node. To use inter-layer communication, it embeds vertical TSVs connecting adjacent stacked layers, as seen in Figure 4. Each layer's SMA is designed to perform inter-layer communication using vertical TSVs and intra-layer communication using horizontal wireless connections. It is equipped with bi-directional, bi-synchronous vertical channels (BBVC), which include bi-synchronous FIFO for up and down ports, to operate at various speeds.

The TSV bus, also known as a vertical wire bus, joins adjacent stacked layers. Figure 5 illustrates the schematically TSV bus linking adjacent stacked layers.

With the help of the transparent vertical wires, the device can add more bandwidth in the vertical direction as needed. Because vertical links can be added or swapped out as needed, the chip has a higher degree of connectivity heterogeneity.

At the right switch boundaries, a TSV with a diameter of at least 4 μm is put into the chip. The area overhead of each TSV in the suggested model is approximately 64 $\mathrm{m}^2$. Each input and output port for a bidirectional vertical switch contains two $(5 + D_{\mathrm{Width}})$ TSVs, where "2" denotes the number of input and output ports, "5" denotes the number of control signals, and $D_{\mathrm{Width}}$ is the width of the inter-switch data link (in bits). The TSV feature aids in decision-directed inter-layer communication by dynamically shifting the data flow direction using a bi-directional channel. The competition among routers for access to output buffers causes congestion in a typical traditional multi-layer NoC architecture, which frequently results in packet losses. Moreover, due to fixed channel direction during the majority of cycles, the "link utilization" of vertical inter-connects either remains low or is idle in some directions. Hence, the suggested model uses bi-directional, bi-synchronous vertical channels that operate at high frequency to effectively utilize the network and reduce congestion. The bi-directional channel dynamically shifts the channel direction between routers in adjacent layers based on the real-time bandwidth availability (i.e. $\mathrm{SMA}_{\mathrm{Link's}}$ parametric $\mathrm{SMA}_{\mathrm{data}}$). The ODA_DD module, which is described in more detail in the next section, is used to accomplish the feature of assessing link use. Overall, by effectively utilizing the links between the stacked layers, the TSVs connecting them aid in power control optimization.

The following criteria are used to evaluate the Network Resource Manager for power control optimization: Think of a NoC with several cores. Suppose that

**Table 4.** Master routing table.

| $\mathrm{SMA}_{\mathrm{id}}$ | $\mathrm{SMA}_{\mathrm{TS}}$ | $\mathrm{SMA}_{\mathrm{node\_status}}$ | $\mathrm{SMA}_{\mathrm{Link\_util}}$ | $\mathrm{MIHA}^{i}_{\mathrm{id}}$ | $\mathrm{MIHA}^{i}_{\mathrm{node\_status}}$ | $\mathrm{MWR}^{i}_{\mathrm{id}}$ | $\mathrm{MPE}^{i}_{\mathrm{id}}$ |
|---|---|---|---|---|---|---|---|

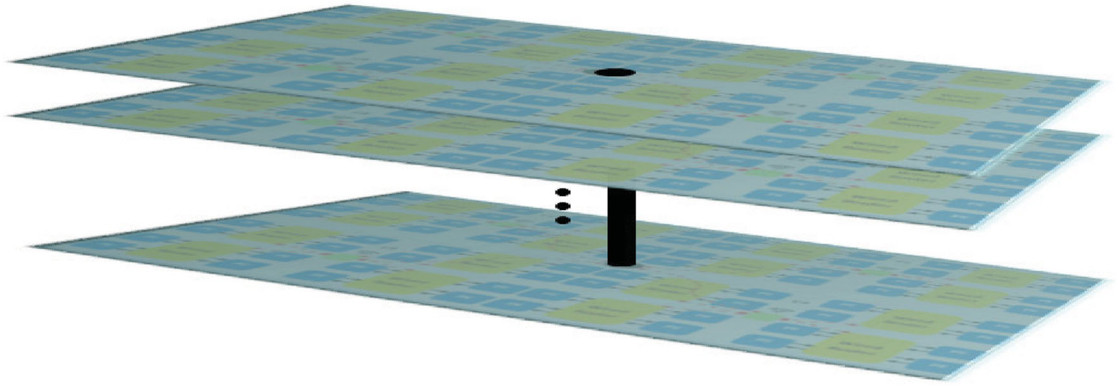

**Figure 4.** Vertical TSV connecting adjacent stacked layers for inter-layer communication.

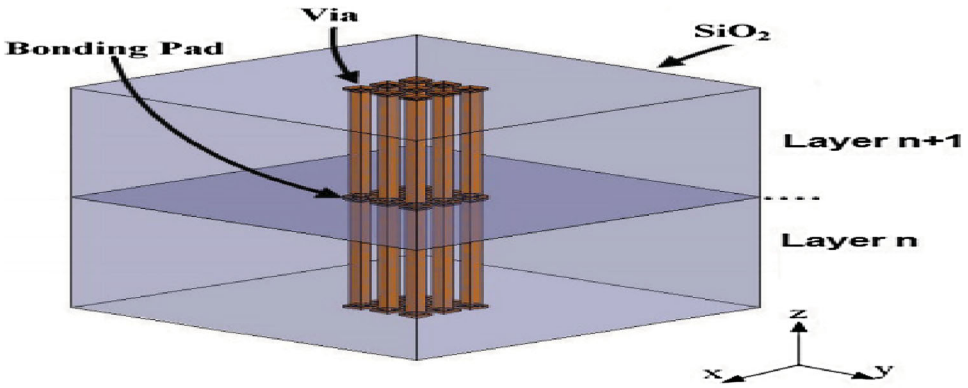

**Figure 5.** TSV bus connecting adjacent stacked layers for decision directed inter-layer communication.

the set $\Omega = 1, 2,$ S contains all of the NoC's cores. Let some of them create and transmit traffic to a few other system cores. Also, allow only one core to serve as the destination for the traffic flow, which should be generated by each of the M cores. The transmission rates of each individual core are represented by a rate vector with the form $r = [r_1, r_2 \dots r_\mathrm{M}]^T$. Additionally, we suppose that the network contains a set $T = 1, 2, \dots$ of links. The route loss gain from the transmitter of link $I$ to the receiver of link $j$ is represented by the element $G_{ij}$ in a path loss matrix $G$ of size $L \times L$ that is dependent on the physical properties of the link. All of the individual transmission powers of the links are contained in the vector $p = [p_1, p_2 \dots p_\mathrm{L}]^T$. A vector $c = [c_1, c_2 \dots c_\mathrm{L}]^T$ is also used to represent the link capacity. A utility function vector $U_i(r_i)$ that is a function of the flow's transmission rate defines each traffic flow $i$. The cost of using the restricted power resources in the NoC system is also represented by a cost function $V_j(p_j)$ for each link $j$ that is dependent on its transmission power $p$ $i$. The Network Utility Maximisation (NUM) formulation is employed for the network.

$$\max_{r,p} = \sum_{i=1}^{M} U_i(r_i) - \sum_{j=1}^{M} V_j(p_j) \tag{4}$$

$$\sum_{i \in z(j)}^{M} r_i \leq c_j \quad \forall\,\mathrm{link}j \tag{5}$$

$$\frac{G_j p_j}{\sum_{k=1,k\neq j}^{L} G_{jk}.p_{k+nj}} \geq \gamma_j \quad \forall\,\mathrm{link}j \tag{6}$$

where $Z(j)$ denotes the set of traffic flows travelling through each link $j$ and $Y(j)$ denotes the desired signal-to-interference-plus-noise ratio (SINR) for link $j$. The entire traffic flow passing from each connection $j$ should not be greater than the link capacity $c\,j$, according to its initial restriction. Please take note that for each link $j$, let $Z$ represent the collection of traffic flows that transit that link $(j)$. Although it can be challenging to determine a wireless link's actual capacity, it is possible to estimate it under certain circumstances using the SINR at each network core.

### Phase III: self-organized centralized service management using SMA's ODA-DD module

As shown in Figure 6, the SMA incorporates an embedded mm-wave wireless transceiver with an ODA_DD module for self-organized interlayer communication. Using centralized service management, the ODA_DD module in SMA optimises inter-layer data transmission.

The following signals make up the physical interface between SMAs for intra-layer communication: $\mathrm{R_x}$: control signal indicating data availability; $\mathrm{data_{in}}$: data to be received; $\mathrm{credit_{out}}$: control signal referring to the

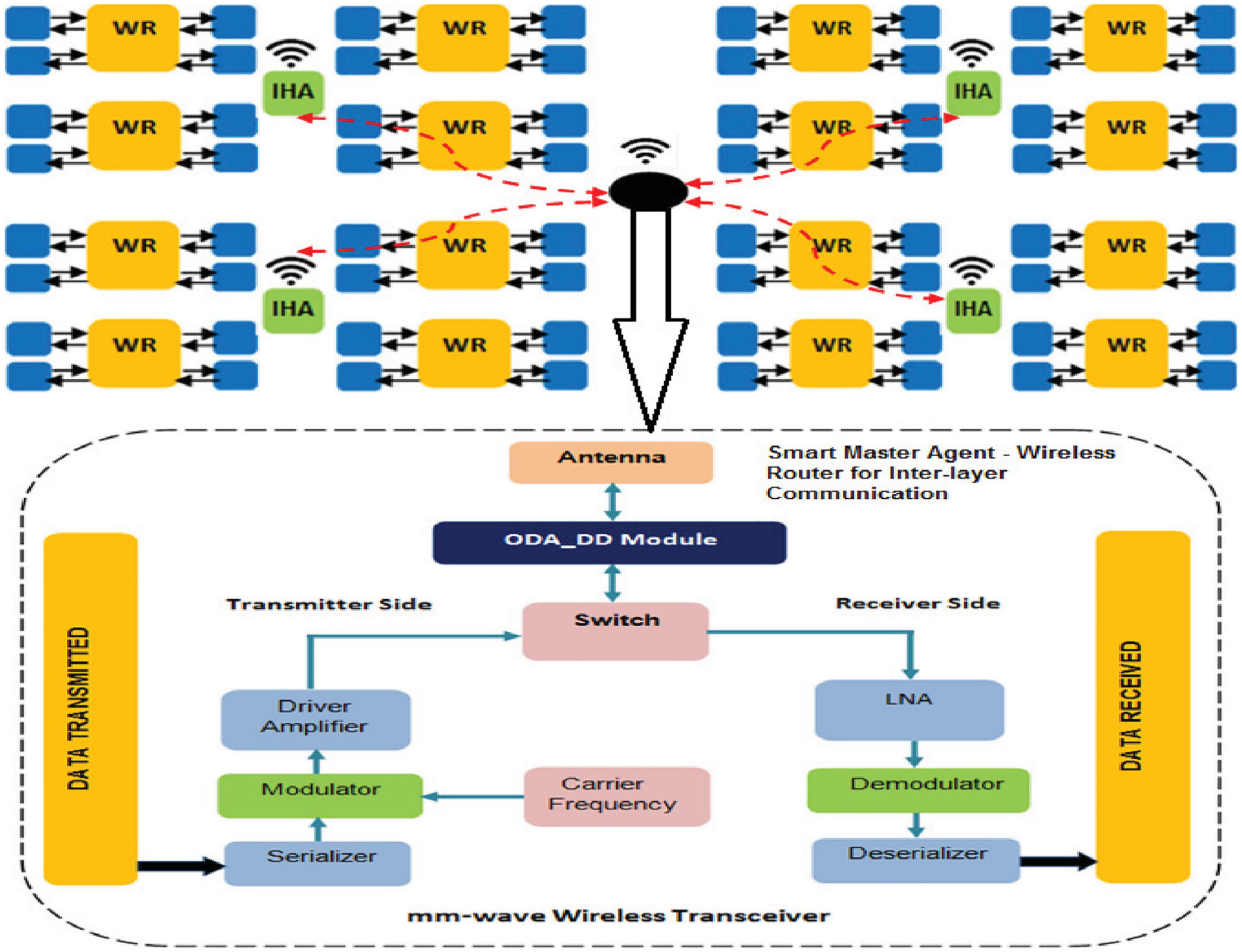

**Figure 6.** SMA mm-Wave wireless transceiver with ODA_DD module for intra and inter-layer communication.

input buffer's space; $T_X$: control signal indicating data availability for transmission; $data_{out}$: data to be sent; $credit_{in}$: control signal indicating space availability in the target buffer; $clk_{write}$ and $clk_{read}$: clock signals. Let's say, for instance, that a SMA (SMA_1) must communicate data to a SMA that is nearby (SMA_2). When this occurs, the neighbouring SMA2 signal asserts the $credit_{out}$ signal, telling the slot to receive data into its buffer. The SMA_1 now asserts a $T_X$ signal timed with the $CLK_{read}$ signal and inserts the data into the $data_{out}$ signal. The transmission is finished when the neighbouring SMA2 receives and stores the data from the input signal. The ODA_DD module enables the network to become self-organized and intelligent for efficient decision-making by self-learning its behavioural features. It has the capacity to adjust on its own to various tasks carried out by the core functional blocks and choose the optimal traffic, allowing it to adjust in accordance with flow needs. Additionally, it regulates the data flow by dynamically altering the channel direction between routers (SMA) in adjacent layers based on the real-time bandwidth requirements (Figure 7).

Since link utilization is a direct indicator of traffic workload, the proposed model uses the link utilization ($SMA_{Link\_util}$) parameter to evaluate the traffic characteristics of vertical channels. As stated in the definition of the SMA Link Utilisation ($SMA_{Link\_util)}$ parameter,

$$SMA_{link\_util} = \frac{\sum_{t=1}^{s} L(t)}{S} \quad 0 \le SMA_{link\_util} \le 1 \qquad (7)$$

where $S$ indicates the window size, $L(t) = 1$ if the traffic passes through the link in the cycle $t$, otherwise $L(t) = 0$.

The $SMA_{Link\_util}$ parameter measures the past communication traffic ($SMA_{Link\_util\_past}$). Using this measurement, the $SMA_{Link\_util}$ of next clock period and the frequency level required for communication is determined. The proposed model makes use of a weighted average smoothing function to ensure the predicted value is reliable. ie.,

$$\begin{aligned} SMA_{Link\_util\_predict} = {} & SMA_{Link\_util\_past} + \alpha \\ & + (SMA_{Link\_util\_actual} \\ & - SMA_{Link\_util\_past}) \end{aligned} \qquad (8)$$

where $SMA_{Link\_util}$ predict denotes the link that will be used in the future (predicted link utilization) and denotes the forecasting weight, $SMA_{Link\_util}$ actual

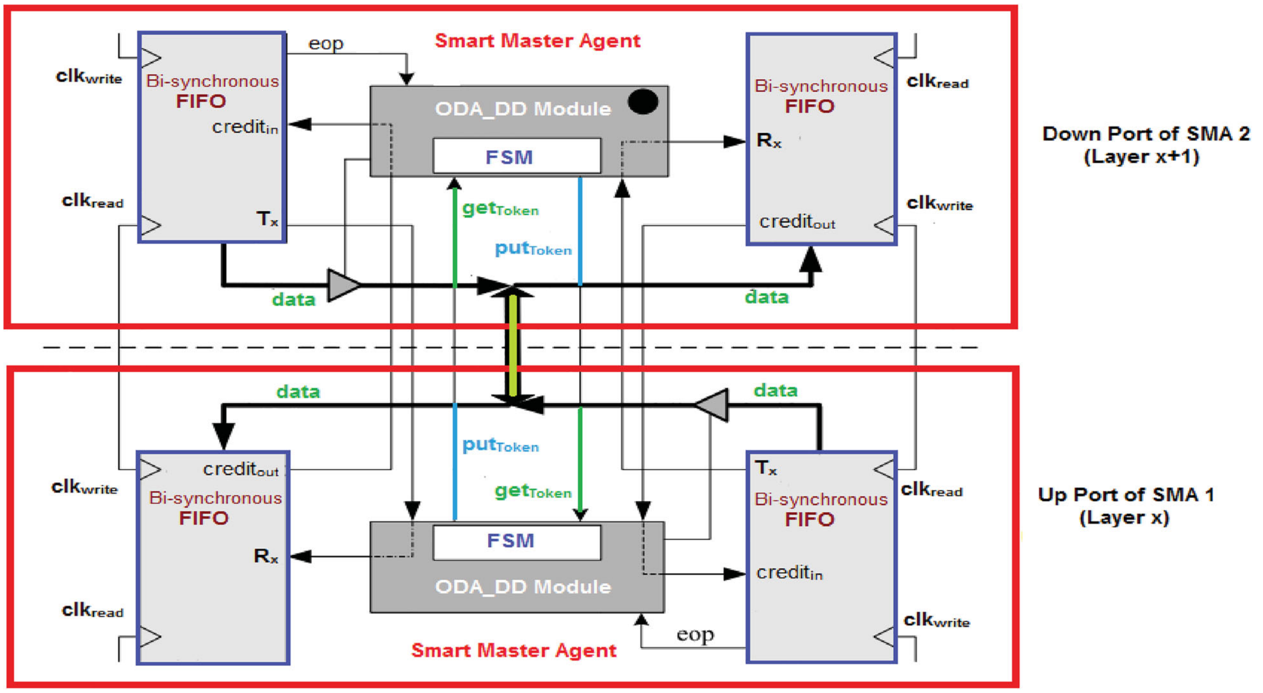

**Figure 7.** Inter-layer data transmission process between SMAs in adjacent layers.

denotes the link that will be used in the current cycle in its actual form, and $SMA_{Link\_util}$ past denotes the link that will be used in the cycle before this one. Since the "weight" function is the only factor that influences prediction accuracy, the weight value is chosen to minimize prediction error. The frequency level is either increased, lowered, or maintained at the same level by the ODA_DD module based on the expected value ($SMA_{Link\_util\_predict}$). It adjusts the frequency level based on long-term traffic profiles to filter out short-term fluctuations brought on by network traffic. Understanding a channel's traffic characteristics is essential for estimating the workload for upcoming communications. In order to determine the characteristics of the traffic, the ODA_DA module uses the link utilization ($SMA_{Link\_util}$) parameter as a network load indicator. The predicted $SMA_{Link\_util}$ parameter, in other words, aids in forecasting future communication workload and the operating frequency needed for data transmission. When there is a lot of traffic, vertical bi-directional bi-synchronous channels perform twice as well as their horizontal equivalents in terms of clock frequency. It provides functionality. The ODA_DD module chooses the frequency level of vertical channels during inter-layer communication to dynamically re-configure and adapt to traffic requirements for effective power management. Figure 8 shows the frequency level forecast using the ODA_DD module.

ODA_DD predicts the frequency level for inter-layer communication using the derived $SMA_{Link\_util}$. Using a counter at each BBVC, the processing unit of the ODA_DA module calculates the actual link utilization ($SMA_{Link\_util\_actual}$) value. In other words, the number of flits passed from the link is counted at each interval. The prediction unit receives as input the $SMA_{Link\_util}$ values actually generated by the computation unit. In addition to the $SMA_{Link\_util\_actual}$ value, the prediction unit also receives the past link utilization ($SMA_{Link\_util\_past}$) value (the link utilization value anticipated from the previous interval) in order to assess the new predicted link utilization ($SMA_{Link\_util\_predict}$) value to be used for the following (S) interval. The Decision Logic Unit (DLU) receives the output value from the prediction unit ($SMA_{Link\_util\_predict}$), which is then stored in the register as $SMA_{Link\_util\_past}$ to be used in the subsequent interval. Simple conditional statements that are equal to the number of frequency levels activated for each timing interval make up the DLU's simple process (H). Every time interval is made up of numerous clock cycles. The DLU calculates the frequency level ($Frequency_{Level}$) to be utilized for inter-layer communication using $SMA_{Link\_util\_predict}$, $SMA_{Link\_util\_past}$, and $Frequency_{Level\_old}$. The ODA_DD module sets the BBVC FIFO mode to either synchronous or asynchronous for effective communication based on the anticipated $Frequency_{Level\_value}$.

Algorithm 1 provides an overview of the procedure for modifying the vertical channel mode depending on frequency level and expected link utilization in the ODA_DD module.

When the frequency level is the same as the intra-layer frequency level, this algorithm's inter-layer and intra-layer communications use the same clock frequencies. The process's non-iterative nature, which makes it straightforward and predictable, reduces the total amount of time needed for execution.

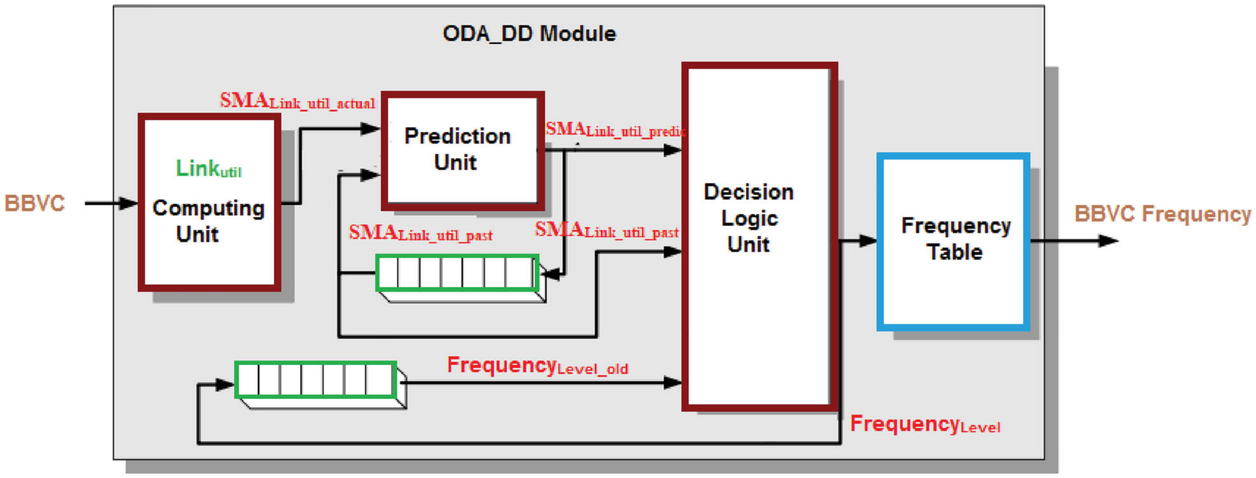

**Figure 8.** Frequency level selection process using ODA_DD module.

**Algorithm 1:** Vertical Channel mode adjustment based on the frequency level in ODA_DD module

```
$SMA_{Link\_util\_predict} = SMA_{Link\_util\_past} + \alpha + (SMA_{Link\_util\_actual} - SMA_{Link\_util\_past})$   //compute link utilization
    if ($SMA_{Link\_util\_predict} < LU_{Threshold\_1}$) then
        $Frequency_{Level} = 1$;
    else if ($SMA_{Link\_util\_predict} > LU_{Threshold\_1}$ && $SMA_{Link\_util\_predict} < LU_{Threshold\_2}$) then
        $Frequency_{Level} = 2$;
        ...
        ...
    else if ($SMA_{Link\_util\_predict} > LU_{Threshold\_n)}$ then
        $Frequency_{Level} = n$;
    end if
        $BBVC\_Frequency_{new} =$ frequency_table ($Frequency_{Level}$);
        $Frequency_{level\_intra} =$ get_frequency_level(intra_layer);
    if ($Frequency_{Level} = = Frequency_{level\_intra}$) then
        $VC_{mode} =$ sync;
    else
        $VC_{mode} =$ bi_sync;
    end if
    $clkT_{past} = clkT_{predict}$
```

## Results and discussion

In this part, an overview of the experimental setup for the DLPM 3DWiNoC model using MATLAB Simulink is provided, as shown in Figure 9(a), and the same figure's packet generation module is used to evaluate the performance in detail (b). The multi-chip architecture gives rise to the formation of a network that is interconnected and is made up of both wired and wireless interconnects. The total number of cores in the system is 256, and there are eight chips in the system. According to rumours, each chip in the system has a whopping 64 cores. Each chip in the system is connected to a central wireless router, the SMA, and the TSVs so that the stacked layers can communicate with one another. This enables inter-layer communication. The on-chip zig-zag antennas that are being considered for use in WIs operate at a frequency of 60 GHz, and they have the potential to provide a bandwidth of 16 GHz. The wireless transceivers that use an OFDMA design achieve a total data rate of one gigabit per second (Gbps) over all channels.

In order to replicate the suggested architecture, a cycle-accurate simulator is employed, and while doing so, the progression of flits via the switches and connections per cycle is tracked. In the case of the data packets, a packet size of 64 flits is taken into consideration, and each flit is responsible for retaining 32 bits. Every simulation is run for a total of 10,000 clock cycles, which enables the transients to stabilize within the first 3,000 cycles and takes into consideration any stalled or routed flits that may have occurred. When a flit needs to be flickered, each wired link in the mesh-based NoC needs one cycle of the clock to do it. Certain on-chip wire line links, on the other hand, require more than one clock cycle for the transmission of a flit. As a result, these on-chip wire line links are pipelined by adding FIFO buffers so that an entire flit can be transferred between two stages in only one clock cycle. Every one of the digital components is driven at a speed of 2.5 GHz, with the clock operating at 1 volt. The simulator accurately models the progression of the flits through each cycle by means of the switches and connections, taking into consideration both flits that have become stuck and flits that have arrived. The performance of mesh-based networks and multichip systems with wired and wireless connectivity is evaluated using the same standard set of metrics, which include throughput, latency, energy,

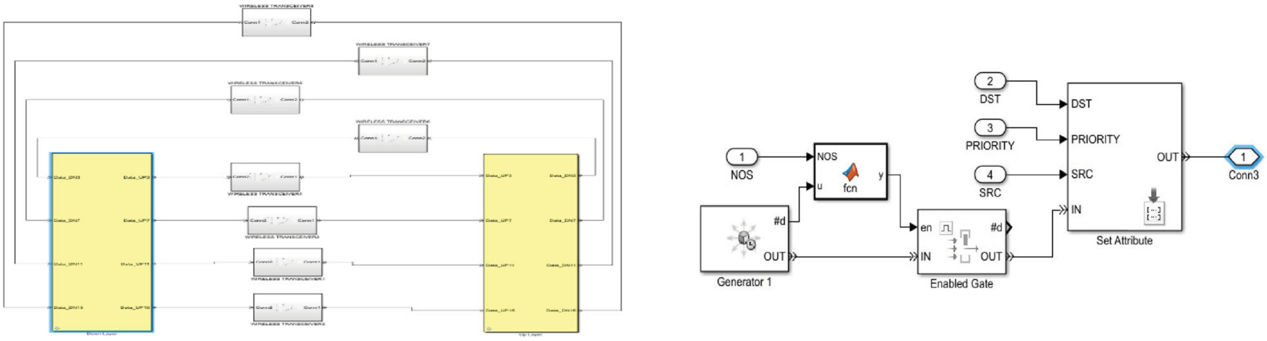

**Figure 9.** (a) DLPM_3DWiNoC Model connecting multiple layers (b): Packet generation module in DLPM_3DWiNoC Model.

power consumption, and area overheads. The primary focus of the comparison study is on benchmark structures such as 2D mesh, 3D mesh, stacked 3D mesh, and ciliated 3D mesh. The kind of traffic that is taken into consideration for evaluation is an auto-injection mechanism [14] that may be found in on-chip modules in a variety of networking and MPEG-2 video applications [17,18]. This method simulates the actual flow of traffic in the environment.

### Throughput

Figure 10 depicts the throughput fluctuation as a function of injection load. As a result of the observation, it is highlighted that the suggested DLPM 3DWiNoC design displays throughput gains over its existing equivalents. While the entire research shows that collision prevention has some limitations, the system's successful routing strategies prevent network saturation before throughput approaches unity by preventing the network from absorbing more traffic than is given.

Bus contention problems limit how much stacked and ciliated 3D mesh-based architecture can improve throughput performance. Even though the DLPM 3DWiNoC architecture uses a bus to connect the different layers, the bus can be expanded and the BBVC service provided by TSVs makes it possible for data transmissions to change on the fly by allowing access to both directions of a channel. In addition, the suggested model offers lower hop counts than strict 2D mesh, 3D mesh, and other existing analogues thanks to the integration of wired and wireless topologies. The ciliated 3D-Mesh topology has somewhat superior throughput characteristics than the stacked 3D-Mesh, 3D-Mesh, and 2D-Mesh topologies. This is due to the fact that it has just 64 inter-switch links, as opposed to 144 in 3D mesh and 112 in 2D mesh. Thus, fewer links typically lead to increased throughput. Also, the DLPM 3DWiNoC architecture's decreased functional IP blocks, improved with a clustering mechanism, are to blame for achieving high throughput without creating congestion during data transmission. In its 2D and 3D mesh cousins, several functioning IP blocks in every switch lower the throughput due to problems with switch contention. It is also obvious that, when compared to the stacked, ciliated 3D mesh and the proposed method, the throughput of the 2D mesh and 3D mesh networks only slightly improves. When compared to the DLPM 3DWiNoC model, the staked and ciliated 3DMesh network does not, however, show the best performance. The findings show that the suggested DLPM 3DWiNoC system has a throughput that is 15%–20% higher than its stacked and ciliated 3DMesh counterparts, making it more reliable and effective than the current method.

### Latency

Figure 11 demonstrates how the amount of packet injection load affects latencies for the designs under discussion.

It is noticed that suggested, stacked, and ciliated 3D designs have less latency than mesh-based 2D and 3D designs. This is because the features of 2D and 3D's hop counts, as well as the routing method, directly affect how long it takes for a flit to traverse the network from source to destination. The development of clusters in each layer with a smart master agent performing centralized service management is another reason why the delay effect in the proposed model is reduced compared to other designs. In addition to reducing the number of hops between nodes, SMA's coordination with its intelligent head agent via wireless link also assists in prioritizing and transmitting more preferred data along the links, preventing packet drops during transmission, and (iii) Inter-layer communication among layers via TSVs allows bi-directional data transmission based on a link's load based on data priority, assisting in delivering emergency data on time to destination and preventing data loss. Whereas the inter-switch wire segments and the switch blocks in 2D and 3D mesh-based, stacked, and ciliated 3D mesh systems form a pipelined communication channel that causes a substantial delay in data transfer. The slowest pipelined stage induced by the longest wire delays accounts for the majority of the overall latency (measured in nanoseconds) in these systems. In other words, the delay of the switch blocks affects the overall latency. Compared to quick vertical

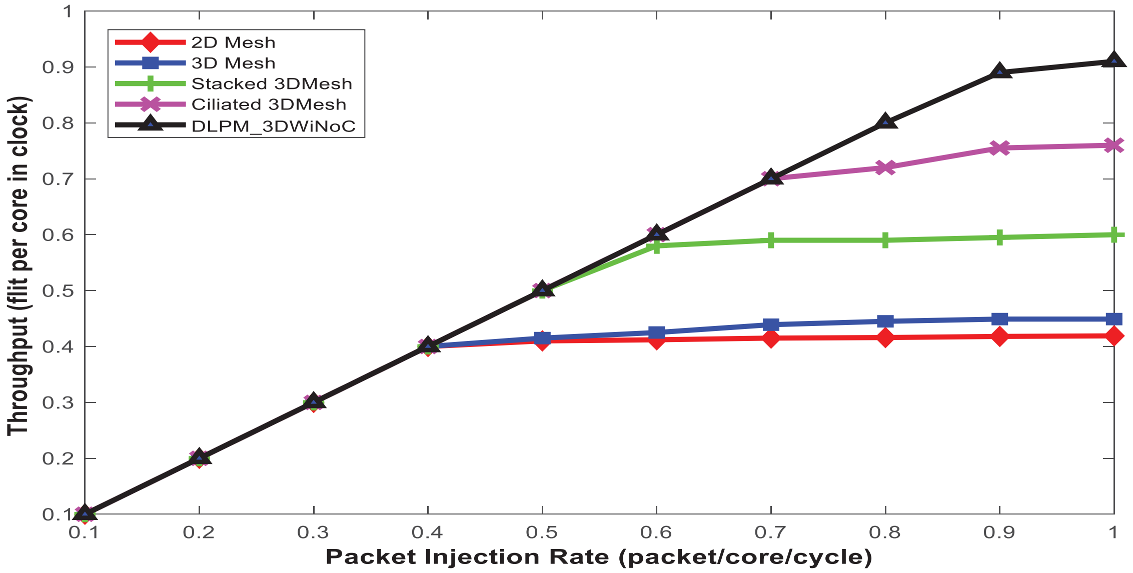

**Figure 10.** Throughput versus packet injection rate.

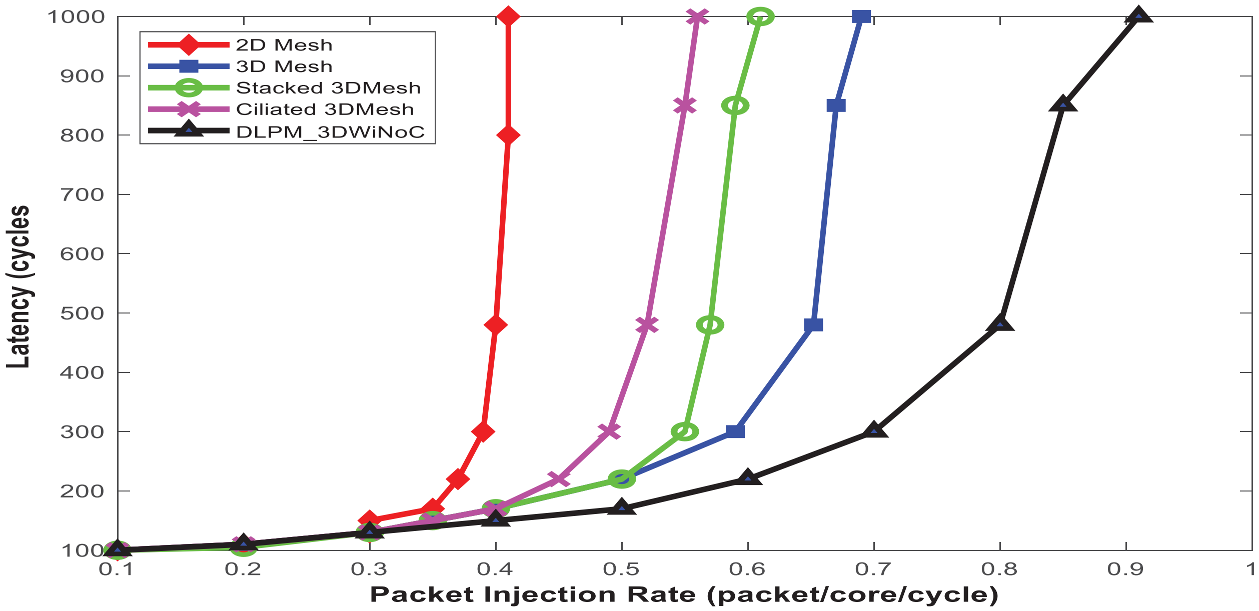

**Figure 11.** Latency versus packet injection load.

links, they are the pipeline's limiting stage. The observation makes it clear that the suggested architectures exhibit the largest performance gain because they successfully shorten horizontal wire runs and add wireless connections between clusters for efficient communication. As a result, when compared to its stacked 3D Mesh and 3D Mesh counterparts, the multi-core, multi-layer DLPM 3DWiNoC model shows regulated latency that is between 12% and 15% lower.

### *Energy dissipation*

Figure 12 shows the packet energy versus several topologies that were taken into account in this investigation. The cost of data transmission is more accurately represented by packet energy. The following variables are the main determinants of energy dissipation: Architecture and packet injection load are both factors. As part of the analysis, these parameters are regarded as independent variables.

The inter-switch segments' and switch blocks' combined energy losses account for the majority of the NoC architecture's energy dissipation. The architecture design affects the number of hops, the packet injection load, and the length of the wires between switches, the number of blocks inside each switch, the energy dissipation factor, and other things. Figure 13 shows how the suggested structures, as well as the 2D, 3D, stacked, ciliated, and 2D architectures, get rid of energy.

The hybrid wired and wireless communication mediums are merged with clustered structures in each tier of the proposed DLPM 3DWiNoC multi-layered architecture. The execution of the design with an IHA in each cluster and the central service management SMA in each tier assures that there is minimum energy loss during packet traversal even when data connection

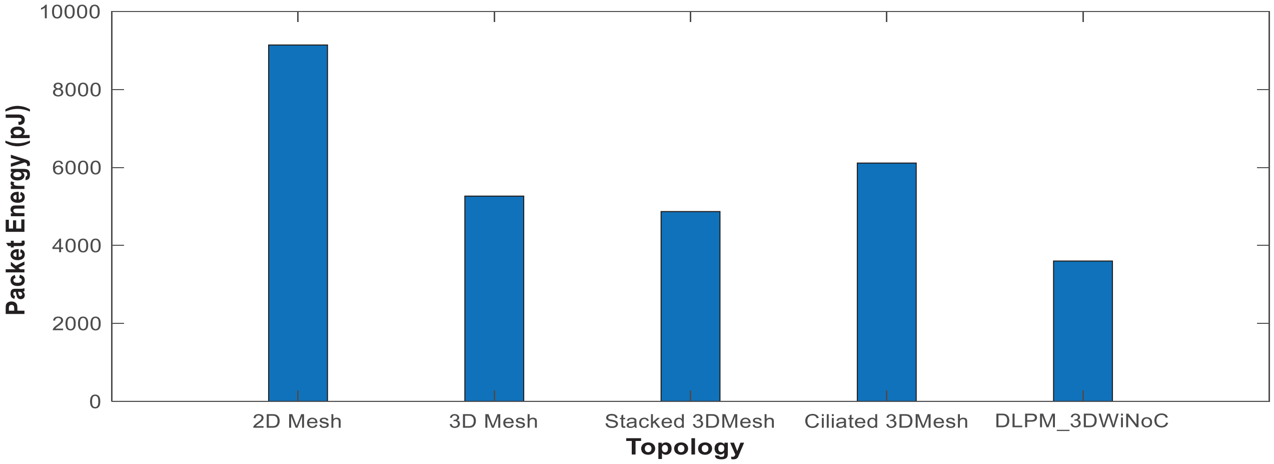

**Figure 12.** Packet energy versus topology.

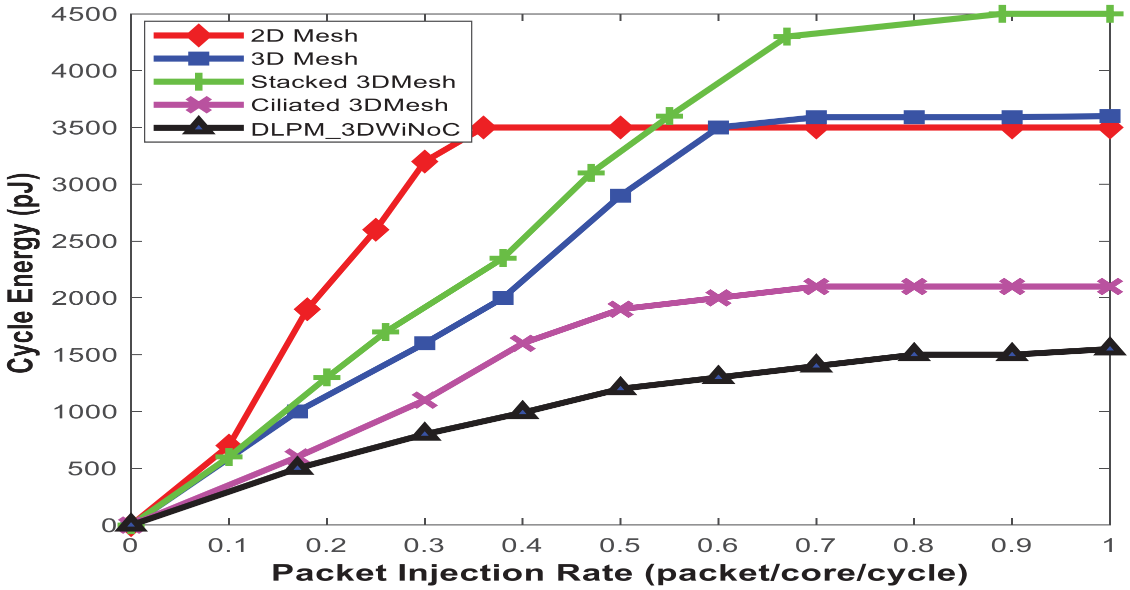

**Figure 13.** Cycle Energy versus packet injection rate.

is triggered among the cores in different layers. The truth is that each core only talks to its WR, which then starts a conversation with its IHA, which in turn creates a wireless connection to the SMA. A source core only needs to make an average of 2–3 hops to convey a data packet to a destination core in the same layer, compared to the significantly higher hop counts required by 2D mesh and 3D mesh. In addition to hop count, the injection load is also crucial to the network's overall energy dissipation. As seen in Figure 12, the proposed DLPM 3DWiNoC technique uses less packet energy than 2D and 3D mesh networks. The existing 2D and 3D mesh approaches actually use more energy during high traffic scenarios because they broadcast more flits at saturation. In contrast, SMA's functionality in the proposed system ensures that transmission flits along links based on its current utility factor. Moreover, flit transmission succeeds even in scenarios with high traffic loads thanks to decision-directed transmission in vertical channels along the line. In addition, we discover that although 3D CITED mesh has fewer hops than 3D mesh, on average it loses more packet energy. When the vertical communication in a 3D ciliated mesh occurs through buses, the capacitive loading on those buses causes a large amount of energy to be lost. Yet, because these 3D designs have the potential for significant energy savings, they have seriously motivated SoC designers to take 3D IC into consideration for future design.

### *Power dissipation*

The overall power consumption of NoC is made up of the power lost through communication lines and routers. The buffers make up around 60% of the leakage power of the NoC among the router parts. The architectural design must take buffers into consideration while attempting to significantly minimize power consumption and maximize network power utilization.

Figure 14 shows the power dissipation of the proposed and existing schemes.

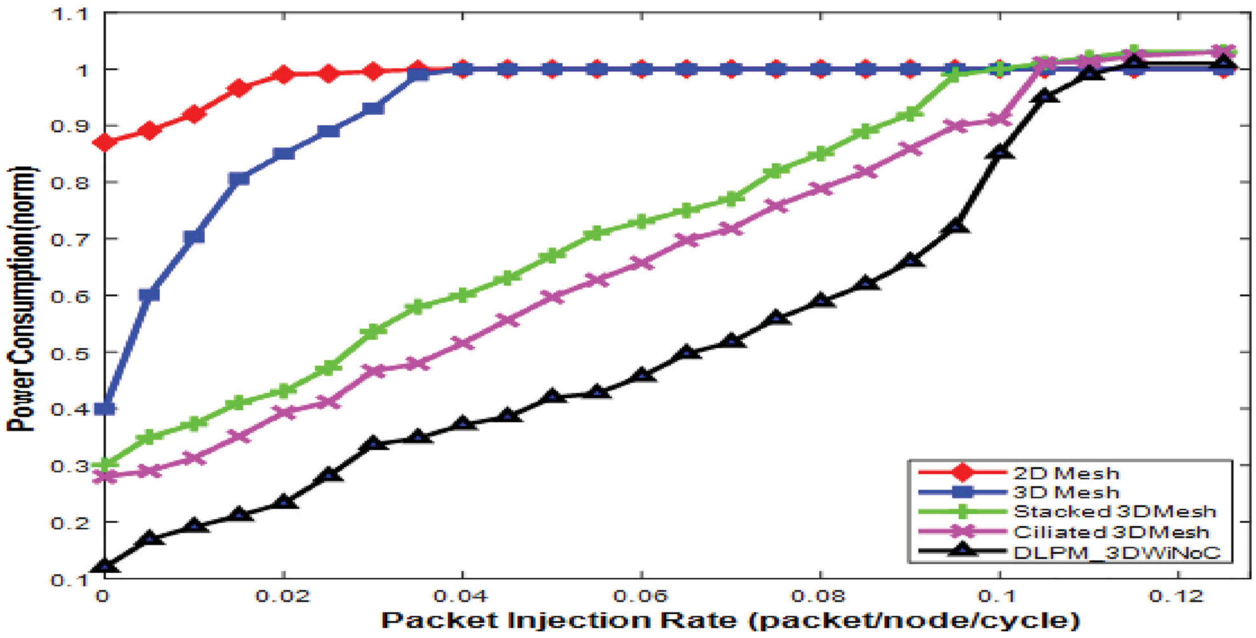

**Figure 14.** Power dissipation versus packet injection rate.

By using a clustering method, the suggested DLPM 3DWiNoC model minimises the number of wired and wireless routers, thereby reducing power consumption in each tier. Additionally, the SMA with TSVs makes it possible for central service management by determining which connection is best to use when transmitting data. In order to successfully transmit flickers across layers, the available vertical channels are chosen in advance. This usually results in successful transmission to the intended location, preventing packet drops, and thereby boosting to some extent the chip's overall performance. The sum of short-circuit, switching, and static power yields the on-chip network's power consumption, which is as follows:

$$P_{\text{consumed}} = P_{\text{short\_circuit}} + P_{\text{switching}} + P_{\text{static\_power}} \\ = \alpha C_{\text{L}} V^{2}{}_{\text{dd}} f + I_{\text{sc}} V_{\text{dd}} + I_{\text{leakage}} V_{\text{dd}} \qquad (9)$$

where the short circuit ($P_{\text{short\_circuit}}$) and switching are parts of "dynamic power consumption" ($P_{\text{switching}}$). The short circuit current ($I_{\text{sc}}$), which develops when the pull-up and pull-down networks in a CMOS circuit are both active at the same time, determines the short circuit power. Whereas the switching power is influenced by the operating voltage ($V_{\text{dd}}$), transition activity factor (TAF), clock frequency ($f$), and load capacitance ($C_{\text{L}}$). Although the static power that dissipates as a result of leakage current ($I_{\text{leakage}}$) is made up of elements like sub-threshold leakage and gate-induced drain leakage (GIDL), Power dissipation in the nanoscale domain is primarily caused by reduced transistor operating voltages and smaller features.

The suggested multi-layer, multi-core DLPM 3DWiNoC paradigm groups the cores in each layer into clusters so they can talk to a common router. The routers, data paths, and buffer architectural components in the suggested model are smaller than those in the current NoC architectural design. Comparatively less total buffer space is taken up overall by a multi-NoC design than by a single NoC. According to the congestion state, the suggested model uses the ODA_DD module to select the suitable IHA and SMA, which then starts the required data transmission along the chosen path. The other equivalent models, such as 2D and 3D mesh, on the other hand, perform data transfer via the open way more frequently when there is a high volume of traffic. The packets wait for the router to be available for processing while they are traversing. This has a large negative impact on performance and a negligible impact on chip power. As a result of increased network latency, there is a power dissipation overhead (the power dissipated due to delay or a drop in flits during transmission). This has a significant impact on the on-chip network's performance, which is essential to the chip's overall performance. According to the discovery in Figure 14, this element causes considerable power dissipation in present methodologies, even at the beginning of the network cycle. It makes intuitive sense that power will rise as more packets move through the network. In order to achieve a suitable trade-off between performance and power consumption throughout the model's design, the efficiency loss caused by power dissipation must be highly compensated to the extent possible. At saturation, it can be shown that all topologies, including 2D mesh, 3D mesh, stacked mesh, and 3D mesh networks, dissipate roughly the same amount of power as the suggested mechanism. It is clear from the observation that the proposed scheme's power dissipation is 15%–20% lower than that of the ciliated and stacked 3DMesh architecture.

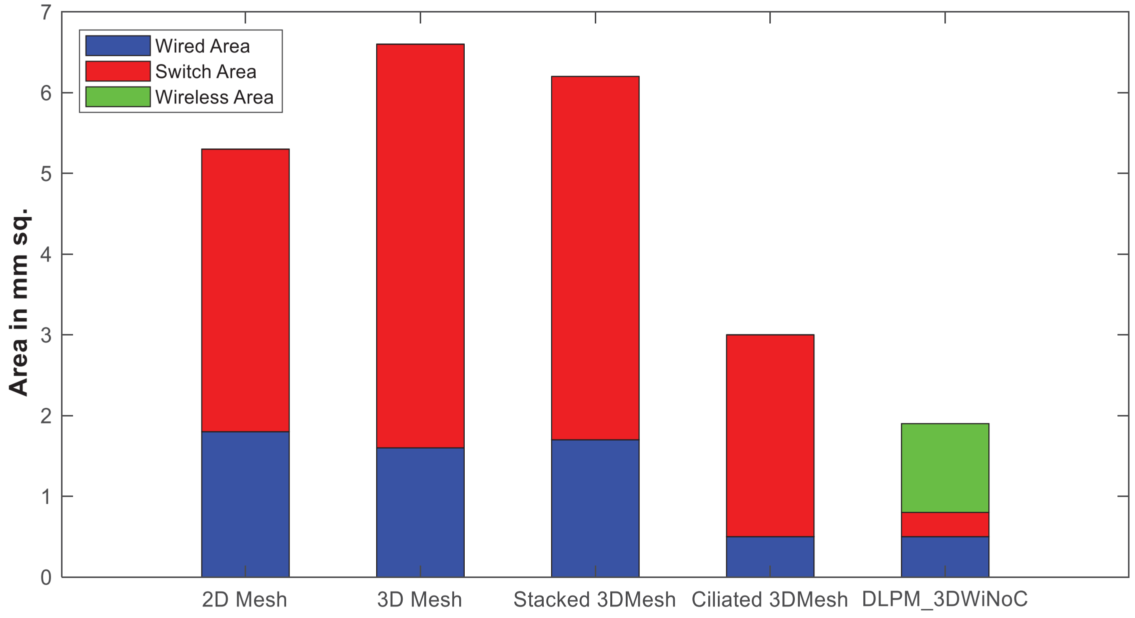

**Figure 15.** Area overheads of wired and wireless architectures.

**Table 5.** Comparison of different architectures.

| | Intra-chip communication | | | Inter-chip communication | |
|---|---|---|---|---|---|
| Architecture | Mesh based NoC | Small-world NoC | Involves wireless links | Serial I/O module | Wireless inter-chip links |
| Mesh + I/O | ✓ | × | × | ✓ | × |
| Small World + I/O | × | ✓ | × | ✓ | × |
| Mesh + CDMA | ✓ | × | ✓ | × | ✓ |
| Small-World + CDMA | × | ✓ | ✓ | × | ✓ |
| Mesh + OFDMA | ✓ | × | ✓ | ✓ | × |

### *Area overheads*

The switch area and wire overhead together make up the area overhead for a NoC. Specifically, the switch area is made up of the total number of switches, the number of ports, and the size of each switch. As mesh-based NoCs have more ports, the area per switch typically grows. While the ciliated design contains fewer switches, reducing the overall switch area, the overall area per switch is increased. When compared to their 2D counterparts, 3D NoCs have more wire overhead due to the horizontal and vertical wirings for intra- and inter-layer communication. Due to the deployment of wireless routers along each layer and a centralized vertical wire deployment through TSVs, the proposed DLPM 3DWiNoC architecture reduces the number of switches and wire length. $L_{\mathrm{ic}} = N_{\mathrm{ip}}$, where $L_{\mathrm{ic}}$ is the chip length on one side and $N_{\mathrm{ip}}$ is the number of IPs or cores in one dimension of the IC, denotes the horizontal wire length. Whereas the horizontal wire length in 2D mesh architectures evaluates to 20 mm (8 or 2.5 mm), it evaluates to 10 mm (4 in 3D mesh structures). This shows that inter-layer wires are responsible for the reduction in wire overhead in 3D mesh systems. The wire area overhead is further decreased in the proposed system, as shown in Figure 15. The DLPM 3DWiNoC architecture mixes wired and wireless communication mediums, prohibiting extended wired connectivity in each layer, which accounts for this. Also, we note that even if the amount of wiring area is decreased in the 3D mesh and stacked mesh NoCs due to an increase in switch overhead, the overall area overhead is considerable, as shown in Figure 15.

The suggested mechanism's overhead is minimal and ideal because it uses wireless technology, fewer routers, and a single centrally located vertical wiring medium for inter-layer communication, making it easier to use and more affordable than its competitors.

*Architectural comparison*: The comparison of five different interconnect architectures are summarized in Table 5 for convenience.

Where,

(i) Mesh + I/O: In this architecture, communication between chips happens through a mesh-based NoC, while communication between chips happens through serial I/O. Each chip's I/O is connected to a single corner switch. This architecture only uses wired connections.

(ii) Small-World + I/O: In this case, interchip communication takes place via a single corner I/O module in each chip, while intrachip communication uses a small-world-based NoC architecture. Its architecture is entirely wired as well.

(iii) Mesh + CDMA: In this architecture, interchip communication exclusively uses the wireless links that connect WIs in different chips, while intrachip communication uses both wireless links and traditional wireline mesh lines.

(iv) Small-World + CDMA: In this case, both wired and wireless linkages are used for intrachip communication, whereas only wireless chip-to-chip links are used for interchip communication.

(v) Mesh + OFDMA: The suggested architecture makes use of TSVs for inter-layer communication between stacked layers and mesh-based NoC with a wireless router for intra-layer communication.

## Conclusion

Despite all of the benefits of 3D NoC, this design paradigm emphasizes a number of difficulties that the NoC research community faces, such as power management and inter-layer modelling issues. The DLPM 3DWiNoC model that is suggested in this work offers a way to optimize power control by utilizing the self-organized, centrally managed service management strategy of the smart master agent, as well as dynamically estimate and efficiently utilize the link during data transmission. Data flow over vertical inter-connects via TSVs is frequently reconfigured by SMA's Optimised Data Communication for Decision Directed Inter-Layer Communication (ODA_DD) module. The simulation findings show that the suggested design demonstrates dramatically feasible performance. Yet, the 3D NoC looks to be a logical progression from the 2D NoC given the development of an efficient 3D integrated circuit fabrication method. The 3D multi-layer NoC architecture does, however, impose a number of unique design and fabrication problems, necessitating more research in this area as future directions.

## Disclosure statement

No potential conflict of interest was reported by the author(s).

## References

[1] Dinesh Kumar TR, Karthikeyan A. Proactive flow control using adaptive beam forming for smart intralayer data communication in wireless network on chip. Automatika. 2023;64(4):689–702. doi:10.1080/00051144.2023.2213927

[2] Hesham S, Rettkowski J, Goehringer D, et al. Survey on real-time networks-on-chip. IEEE Trans Parallel Distrib Syst. 2017;28(5):1500–1517. doi:10.1109/TPDS.2016.2623619

[3] Karkar A, Mak T, Tong K, et al. A survey of emerging interconnects for on-chip efficient multicast and broadcast in many-cores. IEEE Circuits Syst Mag. 2016;16(1):58–72. doi:10.1109/MCAS.2015.2510199

[4] Hwangbo S, Shorey AB, Yoon YK. Millimeter-wave wireless intra-/inter chip communications in 3d integrated circuits using through glass via (TGV) discloaded patch antennas. In: Proceedings of the IEEE Electronic Components and Technology Conference (ECTC). 2016; p. 2507–2512. doi:10.1109/ECTC.2016.278

[5] Liu Y, Pano V, Patron D, et al. Innovative propagation mechanism for inter-chip and intra-chip communication. In: Proceedings of the IEEE Annual Wireless and Microwave Technology Conference (WAMICON), April 2015; p. 1–6. doi:10.1109/WAMICON.2015.7120367

[6] Loi I, Mitra S, Lee TH, et al. A low-overhead fault tolerance scheme for TSV-based 3D network on chip links. In: IEEE/ACM International Conference on Computer-Aided Design, 2008. doi:10.1109/ICCAD.2008.4681638

[7] Hesse R, Nicholls J, Jerger NE. Fine-grained bandwidth adaptivity in networks-on-chip using bidirectional channels. In: IEEE/ACM International Symposium on Networks on Chip, 2012. doi:10.1109/NOCS.2012.23

[8] Kumar S, Jantsch A, Soininen JP, et al. A network on chip architecture and design methodology. In: IEEE Annual Symposium on VLSI, April 2002. doi:10.1109/ISVLSI.2002.1016885

[9] Patti RS. Three-dimensional integrated circuits and the future of system-on-chip designs. Proc IEEE. 2006;94:1214–1224. doi:10.1109/JPROC.2006.873612

[10] Karthikeyan A, Senthil Kumar P. Randomly prioritized buffer-less routing architecture for 3D Network on Chip. Comput Electr Eng. 2017;59:39–50. doi:10.1016/j.compeleceng.2017.03.006

[11] Karthikeyan A, Kumar PS. GALS implementation of randomly prioritized buffer-less routing architecture for 3D NoC. Cluster Comput 2018;21:177–187. doi:10.1007/s10586-017-0979-0

[12] Topol AW, La Tulipe DC, Shi L, et al. Three-Dimensional integrated circuits. IBM J Res Dev. 2006;50(4/5). doi:10.1147/rd.504.0491

[13] Kim J, Nicopoulos C, Park D, et al. A novel dimensionally-decomposed router for on-chip communication in 3D architectures. In: Proc. of International Symposium on Computer Architecture. 2007; p. 138–149. doi:10.1145/1273440.1250680

[14] Feero BS, Pande PP. Networks-on-chip in a three-dimensional environment: a performance evaluation. IEEE Trans Comput. 2009;58:32–45. doi:10.1109/TC.2008.142

[15] Guerrier P, Greiner A. A generic architecture for on-chip packetswitched interconnections. In: Proc. of Design. Automation and Test in Europe Conference. 2000; p. 250–256. doi:10.1145/343647.343776

[16] Rahmani AM, Liljeberg P, Latif K, et al. Congestion aware, fault tolerant and thermally efficient inter-layer communication scheme for hybrid NoC-bus 3D architectures. Networks on Chip (NoCs). Fifth IEEE/ACM International Symposium. Pittsburgh. May 2011; p. 65–72.

[17] Pande PP, Grecu C, Jones M, et al. Performance evaluation and design trade-offs for network-on-chip interconnect architectures. IEEE Trans Comput. 2005;54(8): 1025–1040. doi:10.1109/TC.2005.134

[18] Varatkar GV, Marculescu R. On-chip traffic modeling and synthesis for MPEG-2 video applications. IEEE Trans Very Large Scale Integr (VLSI) Syst. 2000;8(3):335–339. doi:10.1109/TVLSI.2003.820523