International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9s
DOI: https://doi.org/10.17762/ijritcc.v11i9s.7411
***Article Received**: 05 May 2023 **Revised:** 29 June 2023 **Accepted:** 20 July 2023*

Design and Implementation of High QoS 3D-NoC using Modified Double Particle Swarm Optimization on FPGA

Sujata.S.B[1], Anuradha M Sandi[2]
[1]Research Scholar, Department of Electronics and Communication
GNDEC, Bidar,
VTU Belgaum Karnataka, India
e-mail: sujata4sb@gmail.com
[2]Professor, Department of Electronics and Communication
GNDEC, Bidar,
Karnataka, India
e-mail: anu29975@gmail.com

Abstract— One technique to overcome the exponential growth bottleneck is to increase the number of cores on a processor, although having too many cores might cause issues including chip overheating and communication blockage. The problem of the communication bottleneck on the chip is presently effectively resolved by networks-on-chip (NoC). A 3D stack of chips is now possible, thanks to recent developments in IC manufacturing techniques, enabling to reduce of chip area while increasing chip throughput and reducing power consumption. The automated process associated with mapping applications to form three-dimensional NoC architectures is a significant new path in 3D NoC research. This work proposes a 3D NoC partitioning approach that can identify the 3D NoC region that has to be mapped. A double particle swarm optimization (DPSO) inspired algorithmic technique, which may combine the characteristics having neighbourhood search and genetic architectures, also addresses the challenge of a particle swarm algorithm descending into local optimal solutions. Experimental evidence supports the claim that this hybrid optimization algorithm based on Double Particle Swarm Optimisation outperforms the conventional heuristic technique in terms of output rate and loss in energy. The findings demonstrate that in a network of the same size, the newly introduced router delivers the lowest loss on the longest path. Three factors, namely energy, latency or delay, and throughput, are compared between the suggested 3D mesh ONoC and its 2D version. When comparing power consumption between 3D ONoC and its electronic and 2D equivalents, which both have 512 IP cores, it may save roughly 79.9% of the energy used by the electronic counterpart and 24.3% of the energy used by the latter. The network efficiency of the 3D mesh ONoC is simulated by DPSO in a variety of configurations. The outcomes also demonstrate an increase in performance over the 2D ONoC. As a flexible communication solution, Network-On-Chips (NoCs) have been frequently employed in the development of multiprocessor system-on-chips (MPSoCs). By outsourcing their communication activities, NoCs permit on-chip Intellectual Property (IP) cores to communicate with one another and function at a better level. The important components in assigning application duties, distributing the work to the IPs, and coordinating communication among them are mapping and scheduling methods. This study aims to present an entirely advanced form of research in the area of 3D NoC mapping and scheduling applications, grouping the results according to various parameters and offering several suggestions for further research.

Keywords- Router IP, MPSoC, DPSO, 3D-NoC and FPGA.

I. INTRODUCTION

Today's technical advancements enable us to implement ever-more sophisticated applications on computing platforms with potent CPUs. Every processor generation improves upon the one before it, which necessitates loading additional logic onto the silicon [1]. However, there exist two problems with this strategy. The initial problem is that we are no longer able to cut down transistors and the memory and logic blocks they contain [2]. The second problem is that the maximum size of a chip has been reached. Thus, modern technologies incorporate a whole system on one chip, known as a System-On-Chip (SoC), and introduce the idea of hybrid bonding. SoC architecture has improved, moving from a single processor to a multiprocessor (MPSoC), to meet the growing application requirements [3]. The reduction in transistor size, which allowed for the integration of billions of transistors on a single chip, enabled this advancement [4]. Tens of Intellectual Properties (IPs), including processor cores, memory modules, and other I/O factors, may now be supported by MPSoCs [5]. By merging numerous homogeneous or heterogeneous processors, connected to form a network, these systems improve performance. The NoC concept has developed to make it possible to combine several embedded cores onto a single piece of hardware. Traditional linkages in SoC and MPSoC-based buses have been proposed to be replaced by NoC, which ensures flexibility in communication and high requisite bandwidth [6]. The 2D-mesh architecture is the most studied and applied NoC topology due to its consistency and simplicity. However, if networks grow in size due to an increase in core count, network performance may suffer significantly [7]. The suggested strategy to improve performance in terms of power savings and overall hop count is the introduction of 3D NoC [8, 9]. By layering many levels on top of one another and linking them vertically using through-silicon vias (TSVs), a 3D NoC-based MPSoC is created. The ability of 3D NoC-based MPSoCs to combine multiple processors and run more programs necessitates the development of effective and precise mapping and scheduling methods. Prior studies [10, 11] covered a variety of mapping strategies aimed at 2D architectures, while [12] authors reviewed security concerns and responses aimed at standard 3D NoC. To the best of our understanding, task scheduling and mapping are yet unexplored territories in 3D NoC.

Following 3D NoC specialization, the mapping and scheduling procedures have the responsibility of integrating the selected application into the chosen architecture, which entails allocating and organizing the application's operations and communications into the components of the structure in a way that maximizes the objectives of the design. Each node in the application's Directed Acyclic Graph (DAG) denotes a specific task (or computational unit). Every task has a weight that signifies the expense of carrying it out. Two steps are required to complete the process. The DAG's tasks are initially organized in order of execution, and it is called Scheduling. The tasks are classified according to the sequence in which they were completed (1, 2, and 3). The mapping, in which the tasks are to be distributed among the cores, comes next. The results of the simulated network's latency, throughput, and transmission bandwidth can be used to precisely estimate the area and power usage for 3D NoC. The final and most important step in building the communication infrastructure, communication method, and assessment framework for NoC is to associate and choose the right arrangement and deployment of application tasks on various cores. This comprises the fourth dimension, mapping and scheduling in 3D NoC, which is very advantageous in lowering communication costs and improving overall performance [13],[14],[15]. In this work, we primarily focused on this research area to evaluate mapping and scheduling approaches in 3D NoC structures. This work aims to give an in-depth understanding of all the mapping and scheduling components, including task graph generation, scheduling, methods of optimization, the establishment of simulation, and performance evaluation measures. The taxonomy we suggest for organizing all of the papers provided in this work is shown sequentially in Figure 2. In contrast to [16], [17], or [18], our work demonstrates mapping methods employing various categories.

We can transmit huge amounts of data on the same chip by stacking chips on top of each other, which requires shorter and more concentrated connections. A 3D architecture requires hundreds or thousands of micrometer-long connections per square millimeter to link two chips face to face. Data may flow from one silicon piece to another as rapidly and efficiently as if the two existed on the same chip because of these tiny, dense connections. Some important methods for creating dense connections, such as micro bumps, 3D integration, or fusing 2.5D and 3D technologies via Through Silicon-Vias (TSVs), are also involved.

II. LITERATURE SURVEY

Integration of 2.5D with 3D: In a 2.5D structure, dies are placed on Silicon Interposer rather than being stacked one on top of the other. Both dies are chip-chipped onto a silicon interposer, and the dies are arranged in a single package according to a single plan. Using TSVs, interposers, and dies are stacked one above the other in a three-dimensional structure. By merging two 3D stacks containing chip-lets with high-density interconnects, Intel uses 2.5D technology in addition to 3D [19]. Additionally, the dies connect an I/O chip-let and high-bandwidth memory to the largest chip-let (base tile), which is where the remainder of the chips are arranged. The basic tile is then covered with computation and cache chip-lets using 3D stacking technology. Between two chips, the method establishes an extensive network of vertical die-to-die connections.

3D Integration: When many device layers are stacked on a single chip, they are connected using TSVs or Microbumps, and the interconnect May or may not be formed using very-fine pitch TSVs. 3D integration is thought to speed accelerate calculations even if one of the devices in the stack is deficient in even a single transistor. Solder bumps are attached to the ends of interconnects starting at the topmost point (face) of a chip in a miniature variant of flip-chip packaging known as micro bumps. Melting of the solder is then done to establish a connection before the chip is turned onto a matching pair of interconnects on the package's substrate. This method requires the first chip to have tiny copper pillars sticking out from its surface. The two chips are then joined facing each other by heating the solder that has been applied to them and creating a solder micro bump coating. TSVs are vertical copper interconnects that extend down a chip's silicon. Since they do not cover the entire length of a wafer, the silicon's back must be made fine till the TSV is visible. The chips are interconnected in 3D stacked chips to ensure interconnects are facing each other. As a result, the TSVs give the stack permission to acquire the data and power. Graph-core connects a power-delivery chip to their AI processor to accomplish 3D integration [20]. Capacitors and TSVs are tightly packed on the power-management chip.

Hybrid Bonding: *Copper Bonding:* Copper pads at the topmost layer of one chip's interconnect stack are linked to copper pads on other chips using hybrid bonding. In hybrid bonding, the pads are contained in little recesses and are encircled within an oxide insulator. At room temperature, when two chips' copper pads are positioned in opposition to one another, the insulator is activated chemically and it binds immediately. In an annealing phase, the copper pads grow and brace the gap to create a low-impedance link. Compared to cutting-edge micro bump-based methods, this method provides substantial bandwidth and power advantages. Using hybrid bonding, AMD's Zen 3 alongside 3D V-Cache [21] acts as a 3D stacked device that links more cache to a powerful processor.

III. PROPOSED 3D-NOC MAPPING AND SCHEDULING

A network-on-chip (NoC) is used to connect the typical mesh of tiles that make up a 3D-MPSoC. Figure 3 demonstrates it. There is a processing core in each tile, and it is linked to a network router. The overall architecture is made up of links that link together routers. The design makes use of packet-mode communication, therefore all data transported across the network is organized into packets, which are then divided into smaller units known as Flits. Compute blocks as well as the L2 cache are further separated into individual core blocks [22]. Vertical linkages are implemented using TSVs, and lateral communication is accomplished via NoC. Three elements have an impact on 3-D NoC architecture: Topologies, router designs, and a 3-D NoC Design.

A. *The topology of proposed 3D NoC*

The topology of a network dictates how the nodes are connected [23]. Packets may pass through one or more intermediate nodes within a multiple-hop topology before arriving at the destination node. Multiple-hop topologies like mesh and torus are frequently employed in 3D NoCs. In [24], the authors offer a NoC of three-layer with a different configuration for every single layer to meet its specific cost-performance needs. Routers are linked using crossbar switches in a single pillar but at different layers. Feero et al. [9] evaluated and analysed 3-D mesh-based topologies (symmetric 3-D mesh, stacked mesh, and ciliated mesh) along with 3-D tree-based architectures in terms of network efficiency and dissipation of energy.

B. *The Router Designs*

Each step in a scalable node network has a router to manage the usage of shared resources [25]. The typical NoC router functions as shown below. The router computes the permissible output port of a packet while simultaneously writing it into a buffer once it has been received. The packet is subsequently given an unoccupied virtual channel within the output channel using the Virtual-Channel (VC) allocation process. After successfully obtaining a VC, the packet arbitrates access to the crossbar ports. If all goes according to plan, the packet can go through the switch traversal and link traversal steps to a neighbouring router. Communication becomes ineffective when these stages are completed in order because there is a large delay on each hop. Additionally, a packet stalls, lengthening the packet delay, if any of the required resources (such as the buffer or crossbar) are not present. Inhomogeneous architectures have recently been proposed to combine 2D and 3D routers in 3D NoCs to improve performance and save production costs with minimum distortion to the modularity [26]. The limited number of 3D routers in inhomogeneous 3D NoCs, however, results in a performance trade-off. As a result, authors in [27] suggested co-designing routing algorithms in addition to router topologies to overcome the shortcomings of using NoCs with heterogeneous 3D SoC. By integrating the method of low-complexity bypassing and adaptive routing, another effective three-stage pipelined adaptive VC network is presented that equalizes the flow of traffic in hybrid NoCs and accomplishes low-latency communication under diverse traffic loads [28].

C. *The Design Methodology*

By altering or combining topologies, routers, or even vertical links, various concepts for the effective 3D NoC design have been put forth. Data packets are transported between distant cores in a 2D NoC, during which the number of hops significantly lengthens the delay. 3-D integration is used to address the connection delay problem. TSV is a popular 3D integration technique, although it takes up a lot of space on the chip and raises costs because of additional fabrication steps and poor yield. TSVs can be effectively replaced by wireless connection that uses capacity coupling with inductive coupling to address these problems. Inductive coupling is employed by Joseph et al. [29] to create application-specific 3D NoC designs that decrease design complexity brought on by channel interference while boosting channel efficiency. Next, we also need to solve routing difficulties and buffer utilization to enhance the system's overall power, efficiency, and reliability associated with these mesh designs. In contrast to current architectures, the architecture introduced by Rahmani et al. in [30] has a better tolerance for single-bus failure. Considering that dynamic power quadratically declines as supply voltage drops, adopting multiple voltages for supply is a widely recognized low-power option. Siozios et al.'s research [31] demonstrates the inefficiency of using a single power source when creating homogenous NoCs and proposes a 3D NoC design with numerous supply voltages. Furthermore, an advanced mapping algorithm is described that makes it easier to map applications onto such NoCs. The applications are aggregated into groups, and these groups are then mapped into the suggested architecture. Due to the ongoing scaling of process technology, performance and usage of energy are top priorities in today's Systems-On-Chip (SoC) designs. Design engineers may incorporate as many IPs as they can on a chip due to scaling, which enables them to multitask and satisfy customer demands.

However, it results in decreased connection performance, and scaling problems are now a barrier to communication across multiprocessor systems on chips. As a result, the authors of [32] suggest a four-step process for designing and scaling on-chip interconnect designs. The first step is to determine the IP cores' dimensions as well as their likely mapping to the router. The approximate number of IP cores in the MPSoCs overall after clustering, as projected onto the topology, is the core size. The number of generations is determined by the total amount of networks in the topology construction, which is equivalent to the total amount of mapping possibilities. The calculation of the fitness function, resulting in the aggregate energy consumption and communication costs for mapping IP cores to routers, is the second phase. The local and global best cores are chosen in the third stage. Each core offers the most effective, which is the combination of core locations that, among all combinations of that core, has the least EC and CC values. The IP cores with the least network energy and CC values constitute transmission's top performers globally. The evolution over many generations leading to the introduction of a swap operator and a swap sequence whereby the local best and global best are maintained leads to the identification of the ideal solution for the least network energy and CC-based mapping in the final stage. When NoC architecture is extended to three dimensions, the advantages of both strategies are combined, resulting in improved communication performance, increased scalability, and reduced power consumption. The final one is caused by the interconnect capacitance and decreased wire length. At higher integration levels, a crucial contradiction becomes worse despite all these advantages. As device density and power density rise, thermal management, and the necessary cooling options get harder to implement. For typical 3D NoC-based MPSoCs, we investigate various mapping and scheduling strategies that overcome the previously mentioned drawbacks.

IV. SCHEDULING MODEL IN 3D NOC

An important factor in evaluating an application's performance is the time required for computation and communication. The IP cores' architecture largely affects time spent on computation, but the scheduling of the jobs and the routing protocol also have an impact on communication time. To ensure optimal performance, the order in which tasks are completed must be decided. A scheduling algorithm can be used to solve what is known as the scheduling problem. Making a schedule, or a timetable for anticipated circumstances, is called the act of scheduling. The mapping process answers the "where" question, but scheduling is necessary to address the "when" question. Organizing tasks and communications according to a schedule ensures that communications and task executions on the same resource are mutually exclusive. Task scheduling occurs when a large number of an application's tasks are assigned to a single core. Scheduling is the time required for the completion of tasks and communications, which determines how the sequence in which tasks and transactions within them are to be executed so that time limits (for real-time tasks) are adhered to and certain predefined parameters are optimized considering an application task graph assigned onto a 3D NoC architecture. Scheduling issues might be conceptualized as restricted optimization or constraint satisfaction issues. A scheduling issue can be described as 1. A group of intervals that define the actions, procedures, or operations to be carried out. 2. A list of temporal restrictions, i.e., descriptions of possible relationships between the intervals' beginning and ending times. 3. A group of specific limitations, or a description of the complex interactions occurring over a range of intervals as a result of the current condition and limited resource availability. A scheduling strategy is appropriate if it complies with design specifications including timing limitations, data precedence limitations, and memory size limits. The design objectives, such as lowering energy consumption, enhancing temporal efficiency, and balancing memory utilization, should also be optimized. To achieve the same goals, several academics have used various tactics, such as deterministic and heuristic approaches. Others have also used mathematical models like Integer Linear Programming (ILP), Non-Linear Programming (NLP), and Mixed Integer Linear Programming (MILP). These algorithms are categorized into four scheduling subcategories by us. The four scheduling methods are General Scheduling, Task Scheduling, Test Scheduling, and Allocation-Based Scheduling.

Task Mapping: Task mapping is the process of allocating work to the readily available IP cores. Each node in a 3D NoC is linked to a router by a network interface. So, selecting the specific core to be used may simply be added to the task mapping stage. In a standard NoC, connections between cores are made by sending messages hop-by-hop through the source to their destination. By utilizing effective router-bypass techniques, multi-hop traversal is made possible. To facilitate multi-hop traversal, router data pathways, in particular, can be defined both statically and dynamically. This enables flits to fully avoid intervening router pipelines, leading to extremely low latency performance. The performance improvements can only be fully realized when there is no conflict between the flows that are to be discontinued early. In this case, the related flows must be stopped and buffered at the intermediary routers for arbitration, which, in the worst situation, results in hop-by-hop communication. As a result, authors in [92] introduce an SMT-based contention-free task mapping and scheduling framework that addresses this problem by allowing an application task graph to be statically converted to a multi-core or parallel processing system for non-primitive execution based on 2D and 3D. To increase yield, manufacturers are currently using 3D mesh-based NoCs having partially-filled TSVs or reducing the number of TSVs on-chip [93]. However, the topology of 3D NoCs having partially filled TSVs is asymmetric. To address the issue of task mapping in 3D NoC, Ziaeeziabari et al. [89] present a mapping algorithm with irregular topologies and introduce terms like application graph, topology graph, and grading parameter. The topology graph examines the collection of cores and the channels linking the cores, whereas the application graph addresses the sequence of tasks and the communication among them.

Routing: All IP cores can either be combined on one physical plane or spread across several physical planes on a 3D-IC device, which has numerous physical planes. When used in a 3D-IC, an NoC can be either a 2D-NoC (where routers are built on just one physical layer) or a 3D-NoC (where routers are built in many physical plans). In Figure 6, this is displayed. A router in 3D-NoC could link up with routers on a single physical plane or nearby physical planes [52]. The use of an effective routing algorithm is one of the important design objectives to take into account while developing a 3D-NoC. In 3D NoC, there are various routing algorithms and methods that incorporate both routing and mapping. The pipeline design of the conventional XYZ routing method is the Look Ahead XYZ routing algorithm [37]. The augmentation of high traffic in this manner does not take into account virtual channels. Through the use of four pipeline systems—Buffer Writing (BW), Routing Calculation (RC), Switch Arbitration (SA), and Crossbar Traversal stage (CT)—this approach increases throughput. Each hop must pass through all four phases, which decreases performance and increases flit delay. The pipelining system can be used to gather data from the previous stage, and each stage's activities are dependent on the one before it. Because the RC and SA stages are parallel in the Look Ahead XYZ, this dependence between the two phases is abolished. Many routing methods are thought to be inappropriate for NoC applications that have priority requirements for hotspots or the network [104, 105]. To identify the priority of each channel, earlier algorithms added additional buffers to the busy communication channels to overcome this problem. However, these buffers require greater quantities of space and hence use more energy. Instead of employing a round-robin arbiter, Appathurai et al. [81] advise utilizing a lottery arbiter since it can establish the priority for various requests and does not need a second buffer in the event of excessive communication. The varied latency and throughput within every single layer of a 3D network present another difficulty for routing algorithms. To reduce network performance caused by changing throughput and latency of NoCs, Joseph et al. [27] presented a co-design of routing algorithms and network router designs (two routing algorithms termed "Z+(XY)Z" and "ZXYZ"). Because communication inside a layer is synchronous and communication among levels is not constantly synchronous, both vertical and horizontal interactions are modelled separately. The suggested models offer pertinent data on their capabilities and specify the routing algorithms' settings. We describe two routing algorithms based on two concepts that take packet movement across many layers into account. These methods assert to have little area wastage and surpass the latest developments in both theoretical as well as practical evaluations when used in conjunction with router topologies. The architecture of the connection affects a 3D NoC's performance as well as how much power it uses. Numerous routing methods have included BFT topology to improve the efficacy of NoC topologies. In [106], a low-latency, energy-efficient 3D NoC architecture with a zone-based routing approach was presented. With a sizably large network dimension, it reduces the rising traffic load issue that arises in conventional BFT in particular. Similar to this, a recently suggested table-based routing mechanism is detailed in conjunction with a 3D topological design of a NoC based on the BFT topology [107]. The authors of [52] provide another power-aware mapping that incorporates two techniques. According to Elmiligi et al. [52], Dijkstra's algorithm should be used to determine the shortest path routing, and GA should be used to determine network mapping.

V. PROPOSED DOUBLE PARTICLE SWARM MAPPING (DPSM) ALGORITHMS

The proposed 3D-NoC operations depend on mapping and are the process of a Directed Acyclic Graph (DAG) graph to a NoC router. Depending on the previous division of the NoC resource router, the operation on routers that are to be mapped and waiting to be mapped have been identified using DAG which is the main algorithm and swarm intelligence method for best mapping between routers. The graph has N routers and then the structure of particles as a tuple with N elements $R_k^i =$

$\langle E^i_{k,1}, E^i_{k,2}, E^i_{k,3}, \ldots \ldots, E^i_{k,N} \rangle$, where R^i_k is ith particle in the kth iteration and $E^i_{k,n}$ is nth particle in R^i_k graph router corresponding to the nth resource router.

Algorithm 1: DPSM

Inputs:

NoC (U,L): NoC DAG based topology

GR: Group of resource router

P,Q,R: The no of NoC resource routers in the 3D

f(p,q,r): Resource router with coordinates p,q and r

Congestion: Max particle number iterations ($inter_m$)

Output:
Mapping$^{GR}_{TG}$: Mapping relation between GR and TG

```
Initializes particle positions and speeds in a random range: $\text{cong}_1[0], \text{cong}_2[0]$
Measure the fitness of each particle: $\text{fitness}_1[0], \text{fitness}_2[0]$
Initializes forbidden table: $forbidden_{table}$
for loop iterations<$inter_m$ do
Updates cong fitness with equation (1) and (2)
$\text{fitness}_1[iter] < -\text{fitness}_1[iter - 1]$
$\text{fitness}_2[iter] < -\text{fitness}_2[iter - 1]$
Updates inertia factor based on Equation (2)
$\text{cong}_1[iter] < -\text{cong}_1[iter - 1]$
$\text{cong}_2[iter] < -\text{cong}_2[iter - 1]$
for loop $\text{cong}_1$€$\text{cong}_1[iter]$ do
    if Equation (1) then
 $\text{Cong}_1 < -crossover(\text{cong}_1, random(\ \ ))$
     End if
End loop
If $particles_{diversity} < threshold$ then
$forbidden_{table}$<-$\text{cong}_1[iter][0]$
$\text{cong}_1[iter][0]$ searches for neighbourhood and updates if there is a better solution
Nbest,- $\text{cong}_1[iter][1]$
End if
End if
```

Network Latency: To lower the overall hop count, consumption of power, bandwidth requirements, and latency of the network, a mapping algorithm tries to select a core on the chip that is best suited for each task. Techniques for 3D mapping are significantly impacted by the irrationality of 3D NoCs. As a result, the mapping technique is unable to account for each chip core's unique communication abilities. In a 3D NoC, specific topologically close cores may be farther apart than other topologically far cores. In [89], a network latency reduction and hop count minimization approach called 3D-AMAP is presented for 3D NoCs with reduced TSVs and irregular topologies. The algorithm follows four phases of task mapping: (1) *Application Graph partitioning*, where the average communication volume within the application graph is contrasted with the communication volume among tasks, and communications are classified as High and Low volume communications. The algorithm then moves on to phases two through four of task mapping. (2) *High Volume Multi-Tasking Partitioning*, whereby the task having the highest ranking initially maps on a layer with enough spare cores, followed by mapping of tasks with high communication volume corresponding to an established task, and so on, till all the tasks are mapped. (3) *Low Volume Multi-Task Partitioning*, in which task partitions proceed like (2) after being arranged in decreasing order by the total number of intra-partition communications. (4) *Single Task Mapping*, which assigns tasks to an empty core with a low hop count in descending sequence of communications. Low-volume communications are ignored by the algorithm, which also lowers TSV traffic and, as a result, lowers network latency. The delay/Latency is measured based on the time taken to complete the operations on each router in the NoC system, the delay measured by the data in the routing, and the delay of data transmission in the router and it is shown in Equation (2).

$$L_{bit} = \sum L_{route} + \sum L_{core} + \sum L_{router} ------(2)$$

Where $L_{route}$the delay in the data transmission router is, L_{core} is a delay in the source router and L_{router} is a delay in the data in router.

Power or Energy: The processor's power consumption keeps rising as more and more IP cores are being incorporated into it. Reduced power consumption has grown into an important design factor for 3D NoC since it has significant effects on system performance in 3D NoC. The power and energy consumption of 3D NoC designs must be optimized to fulfil the constantly rising communication and low-power demands. The power consumption of NoC for 1-bit data transmission between any source routers to any destination router is shown in Equation (1).

$$P_{bit} = P_{OR} + P_{Cbit} + P_{Abit} + P_{Ibit} ----- (1)$$

Where P_{Cbit} and P_{Ibit} are power consumed on cache memory and power consumed due to internal routing wires. P_{OR} and P_{Abit} are the power consumption for 1bit data transmission on the router and adjacent router links. The tools needed to effectively execute applications at various stages of the design hierarchy are not available to chip designers. A design approach to low-power 3D-NoCs applications must be developed to get the best performance. Elmiligi et al. [52] utilize GA to identify the 3D-NoC mesh network mapping which works best for a particular application and uses the least amount of power. When compared to a thorough search that took three days and still failed to find the lowest power usage, they assert that they were able to solve the optimization problem in just under four minutes. For a specific application, Dijkstra's Algorithm is employed to determine the shortest path routing, and GA is employed to find the best mapping with the least amount of power consumption. [67] Introduces the Bat Algorithm (BA) for the energy-aware mapping method for 3D NoC. To transition between exploring and profiting at the precise moment, a better control approach is needed, and adequate parameter tuning is required for a more effective search. The majority of the energy used by a NoC is used by the routers, and the amount of energy used does not scale linearly with the number of input ports. The Lottery technique is used by Karthikeyan et al. [82] to decrease the power used by asynchronous 3D NoC by recognizing the different input port priorities and guaranteeing that it replies to the port with the highest priority. Energy-aware mapping is a crucial field for research because of the prevalence of battery-operated devices, which raises awareness of the limited power supply. Even though improved energy reduction strategies will enable ongoing performance improvement, they are rarely the only optimization criteria. As a result, there is a trade-off involving energy usage and other user criteria, such as solution quality and efficiency. It is widely recognized that vertical linkages for communication nodes in a 3D NoC are quicker and less energy-intensive than horizontal ones since they are shorter. Therefore, Nalci et al. [42] minimize the energy usage of the application by making maximum use of vertical links for communication nodes. The dynamic energy approach is used by Wang et al. [86] to minimize the energy consumption that results from data transfer across processing units. After defining particular criteria, it categorizes the application as either communication- or computation-centric. If the application is communication-centric, the high-traffic volume communication edges are organized in vertical dimensions to minimize dynamic energy usage. The authors evaluate the suggested approach in comparison to Temperature Balancing (TB) [115] and Temperature-aware Low power mapping (TL) [116] as well as Branch and Bound (BNB) [114]. The results of the studies demonstrate that their suggested approach uses less computing and communication energy. [63] Lists another algorithm that lowers the energy usage of dynamic communication. It uses the BNB method and is a low-energy mapping algorithm. [35] Suggests a Traffic Equilibrium Mapping technique that reduces energy consumption in the chip of a 3D NoC-Bus mesh architecture.

Throughput: The term "throughput" describes how quickly a computing service or device performs tasks. The quantity of packets that effectively traverse the source and destination pathways is referred to as throughput in 3D NoC [117]. The performance and power use of 3D NOC is influenced by the topology of a network. Bose et al. [107] offer a new 3D topological NoC architecture that is based on the BFT architecture with an effective table-based uniform routing algorithm for 3D NoC to achieve an unheard-of performance boost and optimize throughput. Only when the system's temperature remains under control can output be improved. The throughput is optimized by authors in [78], [55] by observing the thermal restrictions. To improve the performance of 3D multi-core processors, Liao et al. [55] propose an approximated task-assignment method that involves modifying temperature equations to produce an incremental thermal update. The temperature of each core is initially calculated using the incremental update approach. The core having the lowest temperature increase out of the unassigned cores is then given a new arrival task. When all new tasks are allocated to cores, the aforementioned procedure comes to an end. By assigning tasks with greater power to upper-layer cores that are closer to the heat sink, authors in [78] enhance throughput. The authors avoid overheating by avoiding assigning tasks to cores at the layer farthest from the heat sink since these cores disperse heat more slowly than others.

When designing and implementing on-chip networks, Quality-of-Service (QoS) is a crucial system-level need. By including at least two signals indicating priority at a transaction-level interface, one of which transfers information in-band with the transaction and the other of which sends information out-of-band with the transaction, QoS requirements can be accomplished in an on-chip interconnect. The on-chip connection can process the signals to provide the necessary QoS. Additionally, a Network-on-Chip (NoC) can be added to the disclosed embodiments. The Quality of Service (QoS) and cost model for communications in Systems on Chip (SoC) are outlined in this study, and the architecture and design process for Network on Chip (NoC) are developed from these concepts. Four kinds of service have been identified for SoC inter-module communication traffic: Signaling (for inter-module control signals), Real Time (for delay-constrained bit streams), RD/WR (for short data access), and Block-Transfer (for huge data bursts). The proposed Quality-of-Service NoC (QNoC) design method examines the target SoC's communication traffic to determine the QoS requirements for each of the four service classes in terms of delay and throughput. Then, a general network design is changed to build a unique QNoC architecture. While maintaining the necessary QoS, the customization process reduces the network cost (in terms of area and power).

VI. RESULTS AND DISCUSSIONS

The proposed 3D-NoC design is synthesized in Vivado Design Suite 2018.1 environment, its utilization summary in terms of device utilization like delay, power, area, throughput, and slice registers are shown in Table.1, and functions of the same are verified using ModelSim simulator. The simulator models the packet transmission delays at each clock cycle of frequency 100MHz and the bandwidth ratio between up and down directions is 100:10. The proposed algorithm to evaluate the performance is compared with Genetic Algorithm (GA) and other experimental results published in recent articles.

Although various scheduling and mapping algorithms assert to optimize of various parameters, the majority of them disregard the growing optimization time. The run-time overhead of the suggested methods can be reduced with more study. The run-time application mapping methodology developed by the authors in [87] can be modified to study the acceptance of changeable process variations and may take into account a service queue model that takes the wait time of a process into account. Similar to [36], faster time frames for task completion can be achieved using 3D mesh architectures without sacrificing thermal limitations. The 3D-NoC technique is more susceptible to overheating and suffers from lower system reliability due to the increase in power density and dissipation of heat. Due to the frequent packet routing that occurs when executing the programs after they have been mapped, the router could be a possible location in this scenario. The router microarchitecture and routing algorithm need to be sufficiently developed to comply with the temperature criteria [110]. To investigate further thermal optimization on 3D NoC architecture, this area requires to be explored more thoroughly and hotspots on NoC routers require being taken into account when doing the mapping. We need to expand our research to include multi-objective optimization, which entails optimizing multiple performance indicators concurrently, as opposed to just taking a single element into account when scheduling objectives. Thus, to further lower the peak temperature and improve performance, DVFS can be used in conjunction with the thermal-aware scheduling algorithm [60]. Security is a crucial issue that has not been covered in previous studies regarding application scheduling and mapping. Developing 3D systems attempts to combine many layers, however since only one layer is linked to the heat sink, it raises the temperature gradient. The 4-layer 3D chip exhibits a temperature standard deviation that is nearly 40 times larger than the 1-layer 2D device [12].

TABLE I. PROPOSED 3D NOC PARAMETERS BETWEEN PROPOSED AND EXISTING

NoC parameters	**Table Column Head**	
	Existing	***Proposed***
Size of flit	256bits	256bits
Router latency	2 clocks, router 1 cycle	4 clocks, router 1 cycle
Buffer Depth	6 flits for 8x8 NoC and 3 flits for 5x5 NoC	7 flits for 8x8 NoC and 4 flits for 5x5 NoC
Routing algorithm	XYZ routing for 3D NoC and XY routing for 2D NoC	XYZ routing for 3D NoC and XY routing for 2D NoC
Baseline topology	4x4x4	4x4x4

TABLE II. PROPOSED 3D NOC UTILIZATION SUMMARY AND COMPARISON BETWEEN PROPOSED AND EXISTING

NoC parameters	**Configuration of simulator**	
	Existing	***Proposed***
No of cycles run	5244	4709
Throughput	80.88GHz	10.8GHz
Latency	133ns	49.7ns
Peak latency	360	57
Power	14.6mW	11.98mW
Slice Registers	1860	1247

Since the matter has grown in significance for high-performance testing systems, there has been an explosion of research attention in thermal-aware test scheduling. Researchers created various mapping along with scheduling techniques for the 3D NoC test scheduling problem to accommodate testing scenarios. Nevertheless, an ineffective solution to the NOC testing issue delivers tests to cores progressively rather than using the full NOC bandwidth, which greatly extends test time as well as expenses [133]. To increase performance in terms of temperature control and overall testing time for 3D NoC systems, we can use splitting of test schedules with combining of tests. The processing power of integrated circuits is increasing, and this intense transistor damaging multi-processor system-on-chips makes communication a more complicated and expensive asset. Network-on-chip (NoC) was developed as a flexible and expanding communication architecture for large SoCs to control communication in such a complicated environment. It addressed several on-chip connectivity issues and enhanced efficiency. Additionally, it offered an outstanding power trade-off for big SoC designs. Routing is crucial to the effectiveness of NoCs. Blockades in the network are inappropriately caused by routing as this can lower performance. Therefore, NoCs have had and continue to have concerns about deadlock-free routing. In this paper, we present PAAD (Partially Adaptive and Deterministic Routing), a unique deadlock-free congestion-aware routing technique. This combines linear and partly adaptive routing techniques that alternate depending on the level of network congestion. Here, a mesh has been separated into various diagonal zones, and various methods are applied to each zone. Deterministic routing is used in each zone when there doesn't exist congestion, while partially adaptive routing is used whenever there is congestion. So it makes use of both and increases total effectiveness. Regarding the delay, throughput, and power performance factors related to different traffic patterns, we analyzed our model with different approaches.

VII. CONCLUSION AND FUTURE SCOPE

Scheduling and mapping algorithms have become increasingly significant as the demand for smaller and faster ICs has grown. This has led to the development of three-dimensional NoC topologies. It is evident that by using efficient mapping and scheduling techniques, the possibilities of 3D NoC design may be completely employed. Since every node in 3D is linked to a router, scheduling, and mapping efficiency are significantly impacted by routing. To satisfy numerous optimization constraints, such as efficiency, communication cost, temperature, consumption of energy, and reliability, this study explores various current mapping and scheduling strategies from an algorithmic viewpoint for typical 3D mesh-based NoCs. Based on the optimization goals, simulators, and benchmarks used, as well as the method used for mapping and scheduling, we organized these studies in this paper. In this research, we first provide a 3D NoC module partitioning approach that is adaptable. The node in the neighborhood with the lowest thermal resistance is searched first, followed by the node having the lowest degree of delay, and finally, the node having the lowest amount of throughput. The DPSO algorithm provides a neighborhood search function to find extra optimal solutions in the solution space close to the optimal and suboptimal solutions, decreases particle convergence in the later phases to decrease the likelihood of dropping into local optimal solutions, and enhances the capacity to search in the initial phase founded on the PSO algorithm. Results reveal that when opposed to the current approach, which is based on a 3D NoC scale of 4x4x4, the average delay is decreased by up to 32.4%, the throughput is increased by up to 35.5%, and energy loss is decreased by up to 31.3%. However, as will be addressed here, there are several opportunities and concerns related to optimization, the ability to generalize dependability, multi-core considerations, and the use or integration of approaches with more intelligent procedures.

REFERENCES

[1] M. Safari, Z. Shirmohammadi, N. Rohbani and H. Farbeh, "WiP: Floating XY-YX: An Efficient Thermal Management Routing Algorithm for 3D NoCs," 2018 IEEE 16th Intl Conf on Dependable, Autonomic and Secure Computing, 16th Intl Conf on Pervasive Intelligence and Computing, 4th Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress(DASC/PiCom/DataCom/CyberSciTech), Athens, Greece, 2018, pp. 736-741, doi: 10.1109/DASC/PiCom/DataCom/CyberSciTec.2018.00128.

[2] M. Beheiry, A. Aly, H. Mostafa and A. M. Soliman, "Direct-Elevator: A modified routing algorithm for 3D-NoCs," 2015 27th International Conference on Microelectronics (ICM), Casablanca, Morocco, 2015, pp. 222-225, doi: 10.1109/ICM.2015.7438028.

[3] E. Taheri, A. Patooghy and K. Mohammadi, "XYZ-ZXY: A minimal routing algorithm for dynamic thermal management in 3D NoCs," 2016 24th Iranian Conference on Electrical Engineering (ICEE), Shiraz, Iran, 2016, pp. 1539-1544, doi: 10.1109/IranianCEE.2016.7585766.

[4] E. Taheri, A. Patooghy and K. Mohammadi, "Cool elevator: A thermal-aware routing algorithm for partially connected 3D NoCs," 2016 6th International Conference on Computer and Knowledge Engineering (ICCKE), Mashhad, Iran, 2016, pp. 111-116, doi: 10.1109/ICCKE.2016.7802125.

[5] R. Salamat, M. Khayambashi, M. Ebrahimi and N. Bagherzadeh, "LEAD: An Adaptive 3D-NoC Routing Algorithm with Queuing-Theory Based Analytical Verification," in IEEE Transactions on Computers, vol. 67, no. 8, pp. 1153-1166, 1 Aug. 2018, doi: 10.1109/TC.2018.2801298.

[6] K. -C. Chen, S. -Y. Lin, H. -S. Hung and A. -Y. A. Wu, "Topology-Aware Adaptive Routing for Nonstationary Irregular Mesh in Throttled 3D NoC Systems," in IEEE Transactions on Parallel and Distributed Systems, vol. 24, no. 10, pp. 2109-2120, Oct. 2013, doi: 10.1109/TPDS.2012.291.

[7] Yadav , A. K. ., Bhaskar Ch. , V. ., M., N. ., Raja, J. E. ., S. Pund, S. ., & Kumari , A. . (2023). A Secure Multi-Path Communication through Dynamic Path Identifiers to Prevent Denial-of-Service Flooding Attacks. International Journal of Intelligent Systems and Applications in Engineering, 11(3s), 22–28. Retrieved from https://ijisae.org/index.php/IJISAE/article/view/2527

[8] F. Dubois, A. Sheibanyrad, F. Pétrot and M. Bahmani, "Elevator-First: A Deadlock-Free Distributed Routing Algorithm for Vertically Partially Connected 3D-NoCs," in IEEE Transactions on Computers, vol. 62, no. 3, pp. 609-615, March 2013, doi: 10.1109/TC.2011.239.

[9] Z. Ghaderi, A. Alqahtani and N. Bagherzadeh, "AROMa: Aging-Aware Deadlock-Free Adaptive Routing Algorithm and Online Monitoring in 3D NoCs," in IEEE Transactions on Parallel and Distributed Systems, vol. 29, no. 4, pp. 772-788, 1 April 2018, doi: 10.1109/TPDS.2017.2780173.

[10] K. N. Dang, A. B. Ahmed, Y. Okuyama and A. B. Abdallah, "Scalable Design Methodology and Online Algorithm for TSV-Cluster Defects Recovery in Highly Reliable 3D-NoC Systems," in IEEE Transactions on Emerging Topics in Computing, vol. 8,

no. 3, pp. 577-590, 1 July-Sept. 2020, doi: 10.1109/TETC.2017.2762407.

[11] A. Coelho, A. Charif, N. -E. Zergainoh and R. Velazco, "FL-RuNS: A High-Performance and Runtime Reconfigurable Fault-Tolerant Routing Scheme for Partially Connected Three-Dimensional Networks on Chip," in IEEE Transactions on Nanotechnology, vol. 18, pp. 806-818, 2019, doi: 10.1109/TNANO.2019.2931271.

[12] R. Salamat, M. Khayambashi, M. Ebrahimi and N. Bagherzadeh, "A Resilient Routing Algorithm with Formal Reliability Analysis for Partially Connected 3D-NoCs," in IEEE Transactions on Computers, vol. 65, no. 11, pp. 3265-3279, 1 Nov. 2016, doi: 10.1109/TC.2016.2532871.A. Charif, A. Coelho, M. Ebrahimi, N. Bagherzadeh and N. -E. Zergainoh, "First-Last: A Cost-Effective Adaptive Routing Solution for TSV-Based Three-Dimensional Networks-on-Chip," in IEEE Transactions on Computers, vol. 67, no. 10, pp. 1430-1444, 1 Oct. 2018, doi: 10.1109/TC.2018.2822269.

[13] Y. Fu et al., "Optimizing Vertical Link Placement and Congestion Aware Dynamic Elevator Assignment for Partially Connected 3D-NoCs," in IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, vol. 40, no. 10, pp. 1957-1970, Oct. 2021, doi: 10.1109/TCAD.2020.3038338.

[14] R. Dash, A. Majumdar, V. Pangracious, A. K. Turuk and J. L. Risco-Martín, "ATAR: An Adaptive Thermal-Aware Routing Algorithm for 3-D Network-on-Chip Systems," in IEEE Transactions on Components, Packaging and Manufacturing Technology, vol. 8, no. 12, pp. 2122-2129, Dec. 2018, doi: 10.1109/TCPMT.2018.2842102.

[15] V. Y. Raparti, N. Kapadia and S. Pasricha, "ARTEMIS: An Aging-Aware Runtime Application Mapping Framework for 3D NoC-Based Chip Multiprocessors," in IEEE Transactions on Multi-Scale Computing Systems, vol. 3, no. 2, pp. 72-85, 1 April-June 2017, doi: 10.1109/TMSCS.2017.2686856.

[16] E. Taheri, R. G. Kim and M. Nikdast, "AdEle+: An Adaptive Congestion-and-Energy-Aware Elevator Selection for Partially Connected 3D NoCs," in IEEE Transactions on Computers, doi: 10.1109/TC.2023.3248260.

[17] Y. Fu, L. Li, K. Wang and C. Zhang, "Kalman Predictor-Based Proactive Dynamic Thermal Management for 3-D NoC Systems With Noisy Thermal Sensors," in IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, vol. 36, no. 11, pp. 1869-1882, Nov. 2017, doi: 10.1109/TCAD.2017.2661808.

[18] P. Guo et al., "Fault-Tolerant Routing Mechanism in 3D Optical Network-on-Chip Based on Node Reuse," in IEEE Transactions on Parallel and Distributed Systems, vol. 31, no. 3, pp. 547-564, 1 March 2020, doi: 10.1109/TPDS.2019.2939240.

[19] E. Taheri, M. Isakov, A. Patooghy and M. A. Kinsy, "Addressing a New Class of Reliability Threats in 3-D Network-on-Chips," in IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, vol. 39, no. 7, pp. 1358-1371, July 2020, doi: 10.1109/TCAD.2019.2917846.

[20] Gan, Y.; Guo, H.; Zhou, Z. 3D NoC Low-Power Mapping Optimization Based on Improved Genetic Algorithm. *Micromachines* **2021**, *12*, 1217. https://doi.org/10.3390/mi12101217

[21] Ponnan, S., Kumar, T.A., VS, H. *et al.* Congestion aware low power on chip protocols with network on chip with cloud security. *J Cloud Comp* **11**, 41 (2022). https://doi.org/10.1186/s13677-022-00307-4

[22] Seth B, Dalal S, Jaglan V, Le D-N, Mohan S, Srivastava G (2022) Integrating encryption techniques for secure data storage in the cloud. Emerging telecommunication technology, Wiley 2020:1–24. https://doi.org/10.1016/j.matpr.2021.01.864

[23] Kun-Chih (Jimmy) Chen.et.al, "Routing algorithm design for power- and temperature-aware NoCs", Advances in Computers, Elsevier, Volume 124, 2022, Pages 117-150, ISSN 0065-2458, ISBN 9780323856881, https://doi.org/10.1016/bs.adcom.2021.11.012.

[24] Savva S, Tatas K, Kyriacou C. Approximate Priority Hybrid 3DNoC Buffered-Bufferless Router. Micromachines (Basel). 2023 Jan 28;14(2):335. doi: 10.3390/mi14020335. PMID: 36838035; PMCID: PMC9961264.

[25] Sujata S.B., Sandi A.M. Design and analysis of buffer and bufferless routing based NoC for high throughput and low latency communication on FPGA. *Int. J. Pervasive Comput. Commun.* 2022;18:250–265. doi: 10.1108/IJPCC-05-2021-0115.

[26] Pontes M.F., Farias C.R., Schvittz R.B., Butzen P.F., Leomar S.R. Survey on Reliability Estimation in Digital Circuits. *J. Integr. Circuits Syst.* 2021;16:1–11. doi: 10.29292/jics.v16i3.568.