ISSN NO: 9726-001X
Volume 13 Issue 02 2025
JOURNAL OF CURRENT SCIENCE

Performance Of Hybrid Connected Network On Chip(NOC) Router To Improve The Latency And Throughput

Raghavendra. A[1], Prashanth. B[2], Ranjith. M[3], Nandini Priyanka. M[4]
[1,2,3] UG Scholar, Department of ECE, St. Martin's Engineering College, Secunderabad, Telangana, India, 500100
[4]Assistant Professor, Department of ECE, St. Martin's Engineering College, Secunderabad, Telangana, India, 500100
appanaraghavendra11@gmail.com

***Abstract*—**The efficiency of NoC architectures is crucial as VLSI systems demand high-speed data processing. Traditional bus-based designs struggle with scalability, while NoCs improve throughput by up to 50%. However, 2x2 NoC routers face high latency and bandwidth constraints, limiting performance. This project proposes a 4x4 NoC router to enhance latency and throughput by integrating look-ahead bypass routing and parallel virtual switch allocation. These optimizations reduce bottlenecks, improve data flow, and ensure efficient resource utilization, making the design highly effective for high-performance VLSI applications.Index Terms—NoC, routing, latency, throughput, VLSI, data flow**.**

I. INTRODUCTION

Network-on-chip (NoC) is crucial in chip multiprocessors (CMPs), facilitating efficient communication between cores. However, NoCs contribute up to 30% of a chip's power budget, and as chip sizes shrink, this percentage is expected to rise. A NoC router consists of input and output ports, a switching matrix, and a local port for accessing connected IP cores. On-chip routers are essential in System-on-Chip (SoC) designs, where improving latency and throughput enhances performance. Optimized topology, adaptive routing algorithms, and Quality of Service (QoS) mechanisms ensure efficient communication and resource utilization.

II. LITERATURE SURVEY

Lee ,et.al[1] developed the Packet classification method, which is crucial in computer networks to increase network security because of developments in high-speed data communication. To support different network services including Quality of Service (QoS), security, and resource reservation, network packet classification is a crucial network kernel function. It became extremely challenging to classify arriving packets using the traditional packet classification algorithms at a decent pace due to the rapidly increasing size of rulesets and rule fields in current networks. In addition to hardware-based solutions, numerous contemporary software-based classification techniques have been put out to speed up packet classification. In general, it's critical to manage low latency, fast throughput, and higher energy efficiency with minimal memory needs while design a packet classification method. AdiSeshaiah, et.al[2]suggested they propose the use of multi-pole nanoelectromechanical (NEM) relays for routing multi-bit signals within a coarse-grained reconfigurable array (CGRA). They describe a CMOS- compatible multi-pole relay design that can be integrated in 3-D and improves area utilization by 40% over a prior design. They demonstrate a method for placing multiple contacts on a relay that can reduce contact resistance variation by 40× over a circular placement strategy.

Chhabria, et.al [3] implemented Neuromorphic processors aim to emulate the biological principles of the brain to achieve high efficiency with low power consumption. However, the lack of flexibility in most neuromorphic architecture designs results in significant performance loss and inefficient memory usage when mapping various neural network algorithms. This literature proposes SENECA, a digital neuromorphic architecture that balances the trade-offs between flexibility and efficiency using a hierarchical-controlling system. A SENECA core contains two controllers, a flexible controller (RISC-V) and an optimized controller (Loop Buffer). This flexible computational pipeline allows for deploying efficient mapping for various neural networks, on-device learning, algorithms. The hierarchical-controlling system introduced in SENECA makes it one of the most efficient neuromorphic processors, along with a higher level of programmability. David Thomas, et.al[4] developed Neuromorphic event-driven systems emulate the computational mechanisms of the brain through the utilization of spiking neural networks (SNN). Neuromorphic systems serve two primary application domains: simulating neural information processing in neuroscience and acting as accelerators for cognitive computing in engineering applications. A distinguishing characteristic of neuromorphic systems is their asynchronous or event-driven nature, but even event-driven systems require some synchronous time management of the neuron populations to guarantee sufficient time for the proper delivery of spiking messages. In this study, they assess three distinct algorithms proposed for adding a synchronization capability to asynchronous event driven compute systems. Gonzalez, et.al[5] suggested Network-on-Chips (NoC) have demonstrated be a favourable alternative to conventional bus-based communication architectures for interconnecting programming elements (PEs). However, NoCs consist of numerous shared resources such as routers and links leading to traffic contention and hence packet transmission delays. Existing works rely on various mechanisms, e.g., leaky buckets, to regulate the network bandwidth distribution and reduce contention. However, such bandwidth regulation mechanisms rarely use runtime information to decide which PEs can inject packets in the network. Milton, et.al[6] implemented as high-performance computing designs become increasingly complex, the importance of evaluating with simulation also grows. One of the most critical aspects of distributed computing design is the network architecture; different topologies and bandwidths have dramatic impacts on the overall performance of the system and should be

explored to find the optimal design point. This work uses destination, their traversal times may increase. Fan, et.al [12] simulations developed to run in the existing Structural Simulation implemented the growing complexity and diversity of neural Toolkit v12.1.0 software framework to show that for a networks in the fields of autonomous driving and intelligent robots hypothetical test case, more complicated network topologies have have facilitated the research of many-core architectures, which can better overall performance and performance improves with offer sufficient programming flexibility to simultaneously support increased bandwidth, making them worth the additional design multi-DNN parallel inference with different network structures effort and expense. Specifically, the test case HyperX topology is and sizes compared to domain-specific architectures. However, shown to outperform the next best evaluated topology by thirty due to the tight constraints of area and power 236 Turkish Journal percent and is the only topology that did not experience of Computer and Mathematics Education Vol.15 No.1(2024),235-diminishing performance gains with increased bandwidth. 241 Research Article consumption, many-core architectures Brouwerian, et.al [7] developed Domain-specific SoCs (DSSoCs) typically use lightweight scalar cores without vector units and are are an attractive solution for domains with extremely stringent almost unable to meet the high-performance computing needs of power, performance, and area constraints. However, DSSoCs multi-DNN parallel inference. Benabdenbi, et.al [13] developed suffer from two fundamental complexities. On the one hand, their Applications taking advantage of these characteristics, with the many specialized hardware blocks result in complex systems and advent of the Internet, have been rapidly democratized and have thus high development effort. On the other hand, their many had a major societal impact: smartphones and social networks for system knobs expand the complexity of design space, making the example. This constant integration has allowed great progress in search for the optimal design difficult. Thus, to reach prevalence, terms of performance but had also required increased attention to taming such complexities is necessary.

everything related to the quality and reliability of the manufactured circuits, especially for circuits targeting critical Taheri, et.al [8] suggested Vertical die stacking of 3D Networks- applications (e.g., aerospace, automotive, health). Densification on-Chip (3D NoCs) is enabled using inter-layer Through-Silicon- exposes the circuit to more defects, defects that can appear at the Via (TSV) links. However, TSV technology suffers from low time of manufacture or later when the circuit is in its final reliability and high fabrication costs. To mitigate these costs, environment. Li, et.al [14] suggested as one of the challenging Partially Connected 3D NoCs (PC-3DNoCs), which use fewer problems in VLSI physical design, global routing is facing TSV links, have been introduced. Nevertheless, with fewer increasing difficulties, and more and more algorithms attempt to vertical links (a.k.a. elevators), elevator-less routers will have to introduce machine learning-based solutions. While most of these send their traffic to nearby elevators for inter-layer traffic, solutions lack high enough routability and routing efficiency. and increasing the traffic load and congestion at these elevators and the average number of the failed nets is around. Andujar, et.al [15] potentially reducing performance. Krestinskaya, et.al [9] suggested Energy efficiency is a must in today HPC systems. To implemented the amount of data processed in the cloud, the achieve this goal, a holistic design based on the use of power- development of Internet-of-Things (IoT) applications, and aware components should be performed. One of the key growing data privacy concerns force the transition from cloud- components of an HPC system is the high-speed interconnect. In based to edge-based processing. Limited energy and this literature, they compare and evaluate several design options computational resources on edge push the transition from for the interconnection network of an HPC system, including traditional von Neumann architectures to In-memory Computing torus, fat-trees, and dragonflies.

(IMC), especially for machine learning and neural network applications. Network compression techniques are applied to implement a neural network on limited hardware resources. Quantization is one of the most efficient network compression techniques allowing to reduce the memory footprint, latency, and energy consumption. Jiaming, et.al [10] developed the computing-in-memory (CIM) technology effectively addresses the bottleneck of data movement in traditional von-Neumann architecture, especially for deep neural network (DNN) acceleration. However, with the improving performance and parallelism of CIM processing elements (PEs), the substantial latency and power overhead caused by high-density intermediate results transmission has become a new bottleneck in CIM architectures. In this literature, they propose a spatial-designed CIM architecture based on the emerging Monolithic 3D (M3D) technology, and a spatiality-aware DNN mapping method for high-performance CIM systems. Ribot Gonzalez, et. al [11] suggested Network-on-Chips (NoCs) have proven to be a good alternative to traditional bus-based communication architectures to interconnect all programming elements (PEs) in modern Multiprocessor Systems- on-Chips (MPSoC). Wormhole switching with Virtual-Channels (VCs) and deflection-based routing policy are the most used strategies to develop NoCs for real-time systems. Deflection based solutions have shown to be the more suitable option for systems with power and/or area constraints. However, because flits may be deflected to alternative routes when traversing the network toward their

III. PROPOSED SYSTEM

A. Overview

This This research focuses on Network-on-Chip (NoC) routers, essential for efficient data communication in multiprocessor systems. The shared bus architecture, where processing elements (PEs) share a common transmission medium, suffers from scalability issues due to increased bus traffic. In contrast, the fully connected crossbar enables parallel data transmission but requires significant area and power. While the crossbar enhances communication, its complexity limits scalability, affecting suitability for systems with numerous processing elements.

B. Proposed Mythology

In the realm of Network-on-Chip (NoC) architectures, the router stands as a linchpin, fostering communication between numerous cores and enabling swift data transmission. Traditional methods like the shared bus and fully connected crossbar delineate two main architectures, each with its pros and cons. Figure 1 shows the proposed NoC architecture.

ISSN NO: 9726-001X
Volume 13 Issue 02 2025

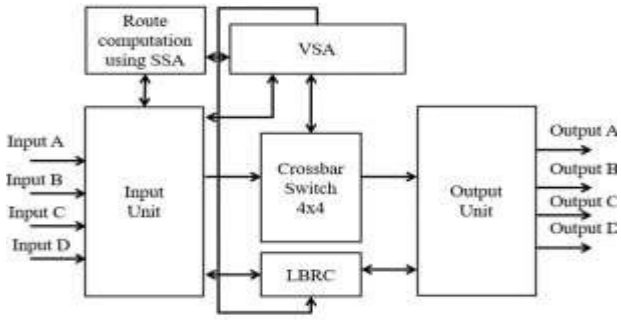

Figure 1. Architecture of the Router.

Step1: -In accordance with the illustration provided, the input data comprises four distinct entities: input A, input B, input C, and input D. Each of these inputs represents a unique stream of data that is directed towards the input unit for processing. Whether it be input A, input B, input C, or input D, each serves as a source of information that contributes to the overall dataset being managed by the system.

Step 2: -The input unit serves as a repository for incoming data, housing a memory bank comprised of multiple registers. Each piece of input data is allocated a dedicated register within this memory bank, ensuring that no data is lost during processing. By utilizing these registers, the input unit effectively manages and stores the required input data without any risk of loss or corruption. This systematic organization allows for efficient data handling, as each input is securely stored in its designated register, ready to be accessed and utilized as needed.

Step 3: -In the computational journey facilitated by the Static Switch Allocator (SSA), a meticulous process unfolds to navigate data from its source to its intended destination. With the presence of four input ports, decisions must be made regarding the most efficient route from the router to the desired output. Whether it's steering data from input A to input B or any other combination, the SSA undertakes the task of path selection, conducting internal assessments to evaluate available options.

Step 4: -In the realm of data management, a critical aspect lies in ensuring the seamless flow of information across various basic pathways. This task is effectively addressed through the utilization of the Virtual Switch Allocator (VSA) method. Operating as a virtual switch allocator, the VSA method plays a pivotal role in overseeing data flow dynamics within the system. In instances where pathways encounter damage or disruption, the VSA swiftly responds by initiating the creation of virtual or temporary paths, thereby safeguarding the continuity of data transmission. Step 5: -In navigating the intricate network of data exchange, the Crossbar switch serves as a critical tool, facilitating the creation of efficient routes for transferring information among various components. As a multi level packet switching network, the Crossbar switch embodies adaptability and versatility, offering a dynamic platform for establishing pathways. Within its architecture, data transfer paths are meticulously crafted, capable of taking on various forms, whether temporary or virtual, to accommodate the evolving demands of data transmission within the system.

Step 6: -When Input A seeks to transmit data to Output B and C simultaneously, it entails a complex data transfer scenario involving one-to-many and many-to-many connections. However, amidst this process, there's a risk of data loss, even with the implementation of virtual switching mechanisms, particularly when dealing with temporary connections across multiple devices. To mitigate such losses effectively, the utilization of LBRC operation becomes imperative. LBRC operation is adept at managing multiple input and output connections concurrently, thereby addressing the intricacies of data transfer within the system. Step 7: -Similar to the input unit, the output unit serves as a crucial component in the data processing system, albeit with a distinct function. While both units handle data transfer, the output unit differs in its approach by exclusively storing output data in memory and registers. Instead of directly processing the data, the output unit focuses on efficiently storing the transmitted information and subsequently transmitting it to output devices for further processing or display. This distinction underscores the specialized role of the output unit in managing data flow within the system, ensuring that output data is securely stored and ready for subsequent processing or presentation.

Advantages from Proposed Methodology

1. Efficient Data Routing with Static and Virtual Switch Allocators
- The Static Switch Allocator (SSA) ensures an optimized selection of paths, reducing congestion and improving packet delivery speed.
- The Virtual Switch Allocator (VSA) dynamically creates virtual paths, ensuring uninterrupted data flow even in case of link failures.
2. Reduced Latency and Improved Throughput
- The crossbar switch and multi-level packet switching optimize data transfer, reducing logic delay (from 14.661 to 4.428 ns) and total delay (from 34.057 to 15.709 ns) compared to the existing 2x2 system.
3. Lower Power Consumption
- The 4x4 router significantly reduces dynamic power consumption (from 9.374W to 0.791W), making it more energy-efficient than the 2x2 router.
4. Support for Multi-Input, Multi-Output Data Transfers
- The LBRC operation enables simultaneous communication from one input to multiple outputs, minimizing data loss and improving efficiency.
5. Better Resource Utilization
- Compared to the existing system, the proposed NoC reduces LUT usage from 492 to 39 while increasing I/O connections from 35 to 69, enabling better hardware optimization.

Applications from Proposed Methodology

1. Multi-Core Processor Architectures
- The 4x4 NoC router efficiently manages communication in chip multiprocessors (CMPs), allowing seamless data exchange between multiple cores.
2. Artificial Intelligence and Deep Learning Hardware

ISSN NO: 9726-001X
Volume 13 Issue 02 2025

- AI accelerators and Neuromorphic Processors use NoC architectures for fast data movement between neurons and processing units, reducing inference latency.

3. High-Performance Computing (HPC) and Supercomputers

- The low-latency routing and hybrid switching improve parallel processing efficiency in supercomputing clusters used in scientific simulations.

4. 3D Integrated Circuits and Embedded Systems

- The methodology's power-efficient routing is beneficial for stacked 3D NoCs, commonly used in mobile processors, automotive control units, and edge AI chips.

5. Cloud Computing and Data Centers

- The proposed NoC's optimized resource utilization and adaptive routing enhance data transmission efficiency in cloud computing infrastructure and server clusters.

Your hybrid-connected 4x4 NoC router significantly improves scalability, power efficiency, and performance compared to traditional 2x2 systems.

IV. RESULTS AND DISCUSSION

Figure 2 presents the results of simulating proposed NoC implementations. Figure 3 provides a summary of the design characteristics of proposed NoC implementations. Figure 4 offers a summary of the power consumption metrics of proposed NoC implementations. It may include information on static power (power consumed when idle) and dynamic power (power consumed during operation), providing insights into the energy efficiency of each design. Figure 5 presumably presents a summary of the time-related metrics of both the existing and proposed NoC implementations. It could include metrics such as total delay, logic delay, net delay, or any other timing characteristics, indicating the latency and performance of each design.

Figure 2. Simulation Outcome

Resource	Estimation	Available	Utilization...
LUT	135	134600	0.10
IO	261	500	52.20

Figure 3. Design Summary

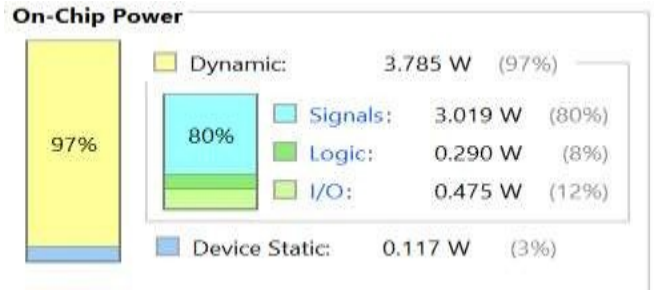

Figure4.PowerSummary

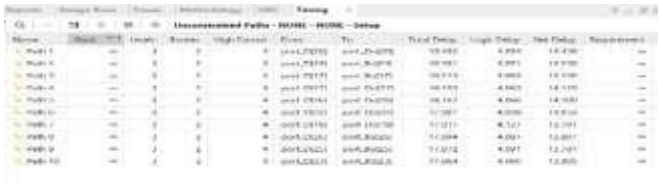

Figure 5. Time Summary.

Table 1 compares the performance of the existing and proposed NoC implementations across various metrics, including resource utilization (e.g., LUTs), I/O count, total power consumption (static and dynamic), logic power, signal power, net delay, logic delay, and total delay. The comparison highlights the differences and improvements achieved by the proposed method compared to the existing one. Figure 6 summarizes the throughput performance of both the existing and proposed NoC implementations. Figure 7 summarizes the latency performance of both the existing and proposed NoC implementations.

Table 1. Existing and Proposed NoC Performance Comparison Table.

Metric	Existing Method	Proposed Method
LUT	492	39
I/O	35	69
Total Power	9.503	0.903
Static Power	0.130	0.130
Dynamic Power	9.374	0.791
Logic Power	4.914	0.076
Signal Power	4.395	0.586
Net Delay	19.407	11.301
Logic Delay	14.661	4.428
Total delay	34.057	15.709

Figure 7. Latency Summary

ISSN NO: 9726-001X
Volume 13 Issue 02 2025

Figure 6. Throughput Summary.

V. CONCLUSION

In conclusion, the hybrid connected NoC router demonstrates a commendable advancement in addressing latency and throughput challenges within integrated circuits. By amalgamating various routing strategies, this innovative approach seeks to strike a balance between efficient data delivery and reduced communication delays. The integration of both deterministic and adaptive routing mechanisms allows the router to adapt dynamically to varying traffic conditions, optimizing latency in diverse scenarios. The significance of packet switching emerges as a focal point in this, highlighting its role in breaking down data into smaller packets for rapid transmission and subsequent reassembly at the destination. The advantages of packet switching over circuit switching, including improved bandwidth, reduced latency, enhanced reliability, fault tolerance, and cost-effectiveness, further emphasize its pivotal role in modern NoC designs.

VI.REFERENCES

100323

[5]Gonzalez, Yilian Ribot, Geoffrey Nelissen, and Eduardo Tovar. "Traffic Injection Regulation Protocol based on free time-slots requests." In 2023 IEEE 29th International Conference on Embedded and Real-Time Computing Systems and Applications (RTCSA), pp. 157-166. IEEE, 2023.

[6]. Milton, Jonathan, and Payman Zarkesh-Ha. "Impacts of Topology and Bandwidth on Distributed Shared Memory Systems." Computers 12, no. 4 (2023): 86.

[71]Boroujerdian, Behzad, Ying Jing, Devashree Tripathy, Amit Kumar, Lavanya Subramanian, Luke Yen, Vincent Lee et al. "FARSI: An early-stage design space exploration framework to tame the domain-specific system-on-chip complexity." ACM Transactions on Embedded Computing Systems 22, no. 2 (2023): 1-35.

[8]. Taheri, Ebadollah, Ryan G. Kim, and Mahdi Nikdast. "AdEle+: An Adaptive Congestion-and-Energy- Aware Elevator Selection for Partially Connected 3D NoCs." IEEE Transactions on Computers (2023).

Krestinskaya, Olga, Li Zhang, and Khaled Nabil Salama. "Towards Efficient In-memory Computing Hardware for Quantized Neural Networks: State-of-the-art, Open Challenges and Perspectives." IEEE Transactions on Nanotechnology (2023).

[10]. Li, Jiaming, Bin Gao, Ruihua Yu, Peng Yao, Jianshi Tang, He Qian, and Huaqiang Wu. "A Spatial-Designed Computing-In-Memory Architecture Based on Monolithic 3D Integration for High-Performance Systems." In Proceedings of the 18th ACM International Symposium on Nanoscale Architectures, pp. 1-6. 2023.

[11]. Ribot González, Yilian, Geoffrey Nelissen, and Eduardo Tovar. "IPDeN 2.0: Real-time NoC with selective flit deflection and buffering." In Proceedings of the 31st International Conference on Real-Time Networks and Systems, pp. 87-98. 2023.

[12]. Fan, Renhao, Yikai Cui, Qilin Chen, Mingyu Wang, Youhui Zhang, Weimin Zheng, and Zhaolin Li. "MAICC: A Lightweight Many-core Architecture with In-Cache Computing for Multi-DNN Parallel Inference." In Proceedings of the 56th Annual IEEE/ACM International Symposium on Microarchitecture, pp. 411-423. 2023.

[13]. Benabdenbi, Mounir. "Contributions to the Test, Fault Tolerance and Approximate Computing of System on a Chip." PhD diss., Université Grenoble Alpes, 2023.

[14]. Li, Chuanyou, Kun Zhang, Yifan Li, Jiangwei Shang, Xinyue Zhang, and Lei Qian. "ANNA: Accelerating Neural Network Accelerator through software-hardware co-design for vertical applications in edge systems." Future Generation Computer Systems 140 (2023): 91-103.

[1] Lee, Youngkwang, Donghyun Han, and Sungho Kang. "TSV Built-In Self-Repair Architecture for Improving the Yield and Reliability of HBM." IEEE Transactions on Very Large-Scale Integration (VLSI) Systems 31, no. 4 (2023): 578-590.

[2]Adiseshaiah, Midde, and Maruvada Sailaja. "A parallel decision-making design for highly speedy packet classification." Microprocessors and Microsystems 99 (2023): 104826.

[3]Chhabria, Vidya A., Chetan Choppali Sudarshan, Sarma Vrudhula, and Sachin S. Sapatnekar. "Towards Sustainable Computing: Assessing the Carbon Footprint of Heterogeneous Systems." arXiv preprint arXiv:2306.09434 (2023).

[4]. Shahsavari, Mahyar, David Thomas, Marcel van Gerven, Andrew Brown, and Wayne Luk. "Advancements in spiking neural network communication and synchronization techniques for event-driven neuromorphic systems." Array 20 (2023):

[15]. Andújar, Francisco J., Salvador Coll, Marina Alonso, Juan-Miguel Martínez, Pedro López, José L. Sánchez, and Francisco J. Alfaro. "Energy efficient HPC network topologies with on/off