Proceedings of CONF-FMCE 2025 Symposium: Semantic Communication for Media Compression and Transmission
DOI: 10.54254/2755-2721/155/2025.GL23423

Network-on-chips: Technological Advances and Future Directions

Yuwen Huang

University of Houston-Dalian Maritime University International Institute, Dalian Maritime University, Dalian, China
hhyywwmant@163.com

Abstract: The network-on-chip (NoC) system is an innovative chip communication architecture that separates communication from traditional chip designs, increasing the scalability of the chip and reducing the burden of the research and development process caused by the increasing complexity. In addition, its centralized communication system significantly shortens chip response time and reduces research and development costs. At present, NoC is a hot research direction in the field of integrated circuits, with broad research potential and application prospects. This paper reviews the advanced literature on network-on-chip systems, summarizes the characteristics and technical trends of NoC to provide researchers in related fields with a concise architecture, and offers some new research directions for future research. The study reveals that the network-on-chip system not only improves the scalability and response speed of the chip but also reduces the research and development cost. Therefore, it has great research potential and application prospects in the field of integrated circuits.

Keywords: network-on-chip (NoC), Multicast routing algorithm, Mapping routing selection strategy, system-on-chip (SoC)

1. Introduction

Moore's Law predicted the rapid increase in the number of transistors on integrated circuits since the invention of integrated circuits in 1958 with a vision that transcended its time. Today, there are more than 100 billion transistors on each chip. While people are pursuing chip miniaturization, they also require chips to maintain high performance to process tasks more efficiently and be applied in a wider range of fields [1]. System-on-chip (SoC) is highly valued by academic and business circles because of its small size and rich functions. However, although SoC has been widely used in many fields such as mobile phones, drones, and the Internet of Things, its research and development process still faces huge challenges in terms of time and economic costs [2].

In addition, as the level of system-level chip integration increases, the complexity of the chip increases dramatically. Cost, performance, power and area together constitute the tasks that engineers need to consider in the process of chip design [2]. To address these issues, network-on-chip (NoC) has emerged as a new communication architecture. Many studies have confirmed that NoC provides an excellent solution for system-level chip communications [3-5]. It meets the current on-chip communication requirements of high efficiency, scalability and low latency [6].

DOI: 10.54254/2755-2721/155/2025.GL23423

This article reviews the research progress in system-level chips in recent years. It aims to provide scholars with a comprehensive understanding of the current research status and potential applications of NoC and to promote the further development of NoC technology.

2. An overview of network-on-chip (NoC)

With the advancement of integration technology, more and more cores can be accommodated on a single chip. The problem is that redundant cores will waste precious space on the chip and will also cause problems for wiring and communication on the chip. For example, the most traditional bus architecture has significant advantages in power consumption and cost, but its scalability is poor and resources are easily competed for, so it is suitable for chips with only a small number of components. The Crossbar architecture performs well in latency and scalability, but its wiring is more difficult [7]. However, the general trend in current chip research is to accommodate as many components as possible on the smallest possible chip, and the above two classic architectures can no longer meet the demand.

The introduction of network-on-chip (NoC) solves these problems well. NoC is a subsystem under SoC, which is specially responsible for communication between various cores. NoC sets up special routers for multiple cores to share, which not only improves the reuse rate of chip space, but also effectively improves the energy utilization rate in complex SoC, reduces losses, and improves the scalability of SoC to make it more flexible [8]. The focus of NoC research is usually on energy consumption, latency, and throughput. Energy consumption is the common goal of the entire SoC, while the lowest possible latency and high throughput are the obligations of NoC as a subsystem used for communication.

As the research on chip architecture deepens, it is discovered that 2D NoC requires long-distance wiring, which occupies a large amount of chip area. People have proposed 3D NoC. 3D NoC increases the utilization rate of limited chip area by adding multiple chip layers in the vertical direction and transmitting data through vertical interconnection to reduce the distance of wires and delay. However, how to stack chip layers without damaging the devices has led to an increase in costs. Overall, 3D NoC has great potential. However, in terms of application, the high cost of 3D NoC poses a considerable challenge to its promotion.

3. Network-on-chip architecture

The foundation of NoC is its hardware, specifically the topology. This architecture lays the foundation for the entire NoC layout, facilitating communication between nodes. It allows engineers to more intuitively grasp the NoC structure, make adjustments, and select appropriate strategies and algorithms. All routing strategies, mapping algorithms and other configurations are proposed based on the NoC architecture.

Traditionally, collective communications are commonly classified into three types: unicast, multicast, and broadcast. Multicast can be further divided into unicast-based multicast routing methods, tree-based multicast routing methods, and path-based multicast routing methods. The unicast-based multicast routing method involves splitting each source node's data into multiple identical packets, which are then transmitted separately to the destination. This method is the simplest and requires minimal equipment, but too many copies of data packets will cause a lot of ineffective power consumption and congestion.

There are two types of tree-based multicast methods: source-specific tree, also called Shortest Path Tree (SPT) and shared tree (RP Tree). The source-specific tree uses the source host as the root of the tree and sends data packets to all receivers, which are leaves, according to the shortest path. In this process, the router will only appear once on a path, thereby ensuring that every path from the root to

DOI: 10.54254/2755-2721/155/2025.GL23423

the leaves is the shortest path. In the shared tree, the source host will first send the data packet to the shared root (RP), and then the RP will send it to the receiver. The advantage of this approach is that it reduces the number of entries in the multicast routing table, reduces the table lookup time and thus reduces latency. However, there is still a situation where the message is copied at the intermediate node and sent to multiple receivers. If one of the branches is blocked, it may cause the entire tree to be blocked [9].

In path-based multicast routing algorithms, the order of receiver transmission needs to be planned in advance, and they need to be listed and placed in the header of the message. Whenever a message enters a router, the router checks whether the router is one of the receivers. If so, a copy of the data is kept and the data is forwarded to the next node, deleting the address of the destination in the header; otherwise, the message is only forwarded to the next node on the path. Many studies have shown that path-based multicast routing algorithms perform better than the other two [10]. Therefore, this paper mainly focuses on path-based multicast routing algorithms.

TLAM, proposed by Yazdanpanah, is a high-performance, low-cost NoC architecture. Since this chip is a NoC architecture proposed for artificial intelligence and machine learning, it has excellent performance in reducing average latency and the number of copies of multicast data packets [11]. The TLAM routing algorithm is based on the 2D Hamiltonian odd-turn model and two-level topology, and is mainly used to solve the deadlock problem in the NoC architecture. By minimizing the number of prohibited paths, the model remains deadlock-free.

The two-layer network topology lays an important foundation for low-cost and low-latency message transmission. Since partitioning the NoC can effectively reduce the average latency and the number of copies of multicast data packets during transmission, thereby reducing congestion, in TLAM, the NoC is divided into four partitions. The first-level link is a common link between routes, and the second-level link is a bidirectional bus that connects the central nodes of the four partitions, that is, the central router to facilitate the rapid transmission of data packets [11]. The biggest innovation of this paper is to set up the routing algorithm by using Hamiltonian algorithm and path-based multicast, and put forward the concept of partitioning NoC to reduce latency. In the simulation part, the above statement was effectively verified. Compared with the 2D and 3D grid HOE-based multicast routing algorithms, both the throughput and average latency have been effectively improved, while having the advantages of low hardware cost and simple routing algorithm.

As chip size shrinks, all types of chips face a common challenge — leakage. As chip power consumption decreases, unnecessary leakage will have a greater impact on the power consumption of the entire chip. This challenge also exists in NoC. Some scholars have proposed to reduce losses by applying power gating [12]. Specifically, it is to specify specific transistors through power gating signals, and physically shut down idle routers to cut off leakage current. However, keeping only some routers working will increase the workload of these routers and increase the response time, and the performance will not be as good as before. At this time, a data packet can be split into multiple sections and distributed to different routers for processing to solve this problem. But the problem that follows is that as the length of the data packet increases, the time to distribute the data packet also increases, which again leads to the problem of increased response time. Wu et al. proposed an architecture called BandExp, which allows routers to "call" the physical links of other routers to expand their own bandwidth and improve processing capabilities [13]. Results show that average packet latency and execution time are significantly improved compared to the state-of-the-art Catnap with only 1.3% of the area used [14].

The architecture provides a framework for NoC, and almost all algorithmic improvements, including routing strategies. It ensures important guarantees for NoC functionality, such as minimizing packet congestion, which can paralyze the entire system. Key research objectives include

DOI: 10.54254/2755-2721/155/2025.GL23423

reducing packet congestion, achieving low energy consumption, high efficiency, and low response time. These are not only the pursuit of NoC, but also the pursuit of the entire integrated circuit. This is an ongoing goal, unlikely to stop until new revolutionary technologies emerge. Currently, research mainly focuses on improving performance by improving layout or algorithms, and research from a physical perspective is rare.

4. Mapping routing algorithm

The mapping algorithm is an iterative process that adjusts parameters in the algorithm and the NoC architecture to achieve overall efficiency improvement [15]. Specifically, the main task is to strategically allocate tasks to different cores in order to pursue lower latency and energy consumption, thereby contributing to the performance of NoC.

Kumar et al. proposed three different mapping algorithms for energy consumption and delay the Horological algorithm, Rotational mapping algorithm and Divide and Conquer Mapping. The Horological algorithm assigns tasks in sequence. As Kumar et al. mentioned, the first task in the queue is assigned to the first core, the second task is assigned to the second core, and so on. When the task on a core is completed, a new task will be assigned to that core. However, since the information is first packaged into data packets and sent to the local router, and then the router finds the adjacent routers to forward it layer by layer, this will cause the access time of the last core to be the longest, reducing the response speed. But in general, this algorithm is particularly effective in promoting load balancing [16]. The rotational mapping algorithm mainly reduces service time by reducing the time spent on mapping tasks. It directly assigns tasks in the queue to the core that is closest to the queue and is idle. Although this algorithm cannot achieve load balancing, it is very effective in reducing latency. However, the biggest problem with this algorithm is that it needs to scan and find idle cores every time it assigns tasks, and time is mainly consumed in this aspect [16].

The two algorithms have clear ideas, simple logic and code. However, Kumar.et al are not satisfied with the above two algorithms and proposed a third algorithm: Divide and Conquer Mapping. Each time a mapping task is executed, this algorithm divides the 2D network topology and assigns the task to the core of the network with the least number of mapping tasks to achieve load balancing. Although this algorithm has significant improvements over the Horological algorithm, it is not as good as the Horological algorithm in terms of service time and latency. In the future, research can focus on reducing latency without paying too much attention to load balancing [16].

Sambangi et al. combined neural networks with reinforcement learning and applied them to 3D NoC, and proposed a mapping method called NeurMap3D [17]. This method is tailored for high-performance computing tasks but is limited by the high computational costs associated with neural networks. Despite its constraints, the potential for this approach is promising as research in neural networks advances.

5. Mapping routing selection strategy

Another key variable that is highly relevant to NoC and also affects data processing is routing strategy. Mapping routing algorithm involves finding the best location or mapping of tasks in a way that the mapping meets certain requirements, such as less energy consumption, less congestion, and less latency, while maintaining a constant bandwidth limit. Mapping routing selection strategy refers to the plan or method used to determine how packets are forwarded or redirected in the network. It involves deciding the best path for packet transmission, such as using a pause schedule to keep packets in a local buffer until they can be sent in the future, thereby improving network reliability and decentralization. In other words, the mapping algorithm determines which core to assign tasks to in order to reduce energy consumption, response time, and other issues. Strategies determine the best

DOI: 10.54254/2755-2721/155/2025.GL23423

path for data packets to be sent to their destinations, especially when exchanging data. It is particularly important to arrange data reasonably to ensure smooth routes.

Trik.et al believe that how the strategy selects the output path is closely related to the overall efficiency of routing, so they proposed a routing strategy, namely "Scored Regional congestion aware and DICA (ScRD)" [18]. First, the authors define a data packet as a non-local request if it must traverse more than two routers. This work is checked by the traffic analyzer every T hours. In the policy algorithm, when the router detects impure local traffic, as long as the non-local traffic rate is greater than 0.4, the router will use the RCA policy; otherwise, the DICA policy will be used. In the case of purely local traffic, ScRD selects an optimal output channel based on congestion levels. In NIRGAM simulations, the results show that the proposed strategy reduces packet latency by 27.10%, increases throughput by 10%, and reduces energy consumption by 6.86% compared to those using either the RCA strategy or the DICA strategy alone. In addition, hardware costs are also reduced.

Routing strategies and other algorithm research share a common goal, achieving low power consumption and low latency. They are different aspects, but they also affect each other. Ultimately, maintaining low response time and minimizing power consumption will remain the primary goals of algorithm research. Although the improvement brought by improving the algorithm may seem marginal, they are significant in the context of integrated circuits, which have entered an era prioritizing low power consumption.

6. Conclusion

This article analyzes the current research progress in the academic field based on the architecture, mapping algorithm and routing strategy of the network-on-chips system, and makes a simple but clear distinction between the three to help scholars who are new to this field or interested in this field to have a simple and superficial understanding. Specifically, the architecture is the skeleton of the NoC, and the algorithm combines with the hardware to pursue performance improvement. The mapping algorithm pursues efficiency at the level of allocating tasks, while the routing strategy determines how to better complete the task.

However, since this article focuses on breadth and makes this article as easy to understand as possible, many more in-depth studies have not been mentioned. Looking ahead, future research can focus on one aspect and analyze it in more detail and professionally. As an effective communication solution for microchips such as system-level chips, NoC has broad commercial and application value. In addition, the pursuit of latency and power consumption is unlimited. As more scholars study NoC, NoC will be applied to a wider range of fields and make greater contributions to chips.

References

[1] *Yin, L., Cheng, R., Ding, J., Jiang, J., Hou, Y., Feng, X., Wen, Y., & He, J. (2024). Two-Dimensional Semic onductors and Transistors for Future Integrated Circuits. ACS Nano, 18(11), 7739–7768. https://doi.org/10.1 021/acsnano.3c10900*

[2] *Cirstea, M., Benkrid, K., Dinu, A., Ghiriti, R., & Petreus, D. (2024). Digital Electronic System-on-Chip Desi gn: Methodologies, Tools, Evolution, and Trends. Micromachines (Basel), 15(2), 247-. https://doi.org/10.3390 /mi15020247*

[3] *Vangal, S. R., Howard, J., Ruhl, G., Dighe, S., Wilson, H., Tschanz, J., ... & Borkar, S. (2008). An 80-tile sub-100-w teraflops processor in 65-nm cmos. IEEE Journal of solid-state circuits, 43(1), 29-41.*

[4] *Wentzlaff, D., Griffin, P., Hoffmann, H., Bao, L., Edwards, B., Ramey, C., ... & Agarwal, A. (2007). On-chip interconnection architecture of the tile processor. IEEE micro, 27(5), 15-31.*

[5] *Marculescu, R., Ogras, U. Y., Peh, L. S., Jerger, N. E., & Hoskote, Y. (2008). Outstanding research problems in NoC design: system, microarchitecture, and circuit perspectives. IEEE Transactions on computer-aided design of integrated circuits and systems, 28(1), 3-21.*

DOI: 10.54254/2755-2721/155/2025.GL23423

[6] Guo, B., Liu, H., & Niu, L. (2024). Network-on-Chip (NoC) Applications for IoT-Enabled Chip Systems: Latest Designs and Modern Applications. International Journal of High Speed Electronics and Systems, 2540027.

[7] Alimi, I. A., Patel, R. K., Aboderin, O., Abdalla, A. M., Gbadamosi, R. A., Muga, N. J., ... & Teixeira, A. L. (2021). Network-on-chip topologies: Potentials, technical challenges, recent advances and research direction. Network-on-Chip-Architecture, Optimization, and Design Explorations.

[8] Zhou, X., Liu, L., Zhu, Z., & Zhou, D. (2015). A routing aggregation for load balancing network-on-chip. Journal of Circuits, Systems and Computers, 24(09), 1550137.

[9] Ebrahimi, M., Daneshtalab, M., Liljeberg, P., & Tenhunen, H. (2010, February). HAMUM-A novel routing protocol for unicast and multicast traffic in MPSoCs. In 2010 18th Euromicro Conference on Parallel, Distributed and Network-based Processing (pp. 525-532). IEEE.

[10] Boppana, R. V., Chalasani, S., & Raghavendra, C. S. (2002). Resource deadlocks and performance of wormhole multicast routing algorithms. IEEE Transactions on Parallel and Distributed Systems, 9(6), 535-549.,

[11] Yazdanpanah, F. (2023). A two-level network-on-chip architecture with multicast support. Journal of Parallel and Distributed Computing, 172, 114-130.

[12] Baharloo, M., Aligholipour, R., Abdollahi, M., & Khonsari, A. (2020). ChangeSUB: a power efficient multiple network-on-chip architecture. Computers & Electrical Engineering, 83, 106578.

[13] Zhou, W., Ouyang, Y., Xu, D., Huang, Z., Liang, H., & Wen, X. (2023). Energy-efficient multiple network-on-chip architecture with bandwidth expansion. IEEE Transactions on Very Large Scale Integration (VLSI) Systems, 31(4), 442-455.

[14] Das, R., Narayanasamy, S., Satpathy, S. K., & Dreslinski, R. G. (2013, June). Catnap: Energy proportional multiple network-on-chip. In Proceedings of the 40th annual international symposium on Computer architecture (pp. 320-331).

[15] Leondes, C. T. (1995). Digital Signal Processing Systems: Implementation techniques. Academic Press.

[16] Kumar, A., Sehgal, V. K., Dhiman, G., Vimal, S., Sharma, A., & Park, S. (2022). Mobile networks-on-chip mapping algorithms for optimization of latency and energy consumption. Mobile Networks and Applications, 1-15.

[17] Ramesh, S., Manna, K., Gogineni, V. C., Chattopadhyay, S., & Mahapatra, S. (2024). Congestion-aware vertical link placement and application mapping onto three-dimensional network-on-chip architectures. IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems.

[18] Trik, M., Akhavan, H., Bidgoli, A. M., Molk, A. M. N. G., Vashani, H., & Mozaffari, S. P. (2023). A new adaptive selection strategy for reducing latency in networks on chip. Integration, 89, 9-24.