Proceedings of CONF-FMCE 2026 Symposium: Smart City and Infrastructure Engineering
DOI: 10.54254/2755-2721/2026.CH37633

Network-on-Chip Technology for Chiplet Integration Architecture

Tianmiao Zhang

Beijing No.80 High School, Beijing, China
Tianmiao1063@outlook.com

Abstract. In today's era, the marginal benefits of transistor scaling are continuously decreasing, and Chiplet heterogeneous integration has become the mainstream path for scaling AI chip computing power. Meanwhile, on-chip interconnect performance is increasingly becoming the key factor limiting system computing power and energy efficiency. This paper focuses on on-chip network technology for Chiplet-integrated AI chip architectures, providing a systematic review and analysis of its technical principles, research status, existing limitations, and development directions. The study first breaks down the on-chip network's dual-layer architecture of 'physical architecture layer - protocol mechanism layer,' clarifying the functional roles of each core hardware unit and protocol mechanism and the complete data transmission logic. Next, from the perspectives of topology, routing algorithms, switching mechanisms, and flow control mechanisms, it summarizes the evolution of on-chip network technology and the customized research results for AI computing features and Chiplet heterogeneous scenarios. It then identifies two core current issues—inter-chiplet interconnect performance bottlenecks and lack of cross-layer collaborative design, as well as insufficient dynamic load adaptation and lagging industry standards — and outlines the key paths for future technology optimization. This research can provide clear insights into the technological development path for researchers in the field and offer theoretical support and direction for the interconnect architecture design and industrial implementation of the next-generation highly integrated AI chips.

Keywords: Network-on-chip, Chiplet integration architecture, AI chips, Interconnect optimization

1. Introduction

With the growing demand for AI computing power, the efficiency shortcomings of traditional CPU architectures are becoming more obvious. The computing unit and the memory unit are physically separate in traditional CPUs' architecture, so data must be moved back and forth between them to perform calculations. In terms of bandwidth, energy, and delay, data transportation is far costlier than computing [1]. The bandwidth density, latency, and power performance of chiplet interconnects are obviously worse than on-chip connections. Data exchange across chiplets adds extra communication overhead. Current interconnection designs have become the limiting factor in further

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2026.CH37633

scaling performance and energy efficiency. The low throughput of silicon interposers renders inter-chiplet communication costly [2].

There are mainly three kinds of needs in multi-core architectures: scalable inter-core connections, cache coherence and memory hierarchy optimization, and resource isolation with coordinated scheduling. General-purpose multicore architectures struggle to handle all workloads. Performance and power of gigascale systems-on-chip is increasingly communication-dominated. Under constrained power budgets, designers must maintain overall system performance while meeting the communication requirements of an increasing number of integrated cores [3]. When the number of on-chip computing cores grows to dozens or even hundreds, traditional interconnect methods like shared buses and crossbar switches have inherent issues like insufficient bandwidth. The bandwidth and energy need for their connectivity networks rise exponentially with the number of cores per multiprocessor, and it is anticipated that traditional on-chip wires will be unable to match these demands [4]. Therefore, developing high-performance on-chip network technology can enable large-scale, high-bandwidth, low-latency, and highly reliable data exchange on chips.

The development of on-chip networks can generally be divided into three stages. The first stage is the theoretical foundation stage, when the scalability limits of traditional shared buses and crossbar switches began to emerge. In 2002, Benini and De Micheli clarified key research aspects like topology, routing algorithms, flow control mechanisms, and network interfaces, building a complete theoretical framework for the field. Research in this stage ensured functionally accurate and dependable operation of interacting system-on-chip components will be handled by on-chip micro networks, which are created using a layered technique [5]. The second stage is the period of rapid technological development and industrial implementation. Chip multi-processors, system-on-chips, and system-in-package design could be revolutionised by combining 3-D integration with scalable interconnects like NoCs [6]. During this time, 3D NoCs became a hot research topic. By using vertical through-silicon vias for short-distance connections between multiple chip layers, it could significantly shorten interconnect lengths and increase bandwidth density. At the same time, NoC research continued to expand toward engineering needs, being able to meet different service requirements for latency and bandwidth. To address signal crosstalk issues in deep submicron processes, fault-tolerant routing, data retransmission, and other mechanisms were continuously improved. By the later part of this stage, NoC technology had become the mainstream solution for interconnecting multi-core chips. The third stage is all about cutting-edge breakthroughs in scenario customization and cross-level integration. For the data flow features in deep learning tasks, researchers have come up with solutions like adaptive topology design and have even used smart algorithms like reinforcement learning to optimize network strategies. Over the course of several generations, chiplet connection standards have rapidly advanced, greatly increasing data speeds, energy efficiency, and application variety [7].

Although NoC technology has quite mature research results in areas like infrastructure, 3D stacking, and low-power design, there are still noticeable shortcomings when it comes to AI chips under the emerging Chiplet integration architecture. Therefore, this paper provides a comprehensive review of on-chip network technology for Chiplet integration architectures, covering basic principles, current research status, existing limitations, and future directions, aiming to offer ideas for developing more efficient and highly integrated AI chips.

2. Principal of NoC

NoC borrow the concept of packet switching from distributed computer networks, restructuring the interconnection of on-chip computing resources from the traditional bus-based shared path into a

DOI: 10.54254/2755-2721/2026.CH37633

modular network communication architecture. The NoC architecture essentially is the on-chip communication infrastructure comprising the physical layer, the data link layer and the network layer of the OSI protocol stack [8]. The physical architecture layer of the NoC is the hardware foundation, and the four main components of this layer are computing resource nodes, routers, network interfaces and physical links. Computing resource nodes act as the communication endpoints for the network and contain CPU cores, GPU compute units, NPU acceleration cores, cache modules, various dedicated IP cores, etc.; they are the sources and consumers of data packets. Routers are the central forwarding components of NoC, and they contain input buffers, crossbar switches, routing computation units, and arbiters to receive and parse data packets, calculate routes, and forward them.

The Microarchitecture Design Directly Impacts Network Latency, Throughput and Power Consumption. Network interfaces are located at the connection points of computing nodes and routers; they perform protocol conversion by packaging communication requests from computing nodes into data packets or converting received packets back into node-recognizable transactions, and handle clock domain crossing. Physical Links are the on-chip interconnection lines connecting neighbouring routers, and their width and driving mode directly affect the transmission bandwidth and power consumption. These hardware units are connected in a certain way to form a network topology. The Topology of a NoC determines its network diameter, node degree and scalability, etc., so it is a fundamental choice in NoC design. The protocol layer provides the rules for packet transmission and resource scheduling in the network, and routing algorithms, switching mechanisms, flow control mechanisms, etc., work together to achieve this goal. The Routing Algorithm selects a path for a packet from the source to the destination, improves load balancing, and should avoid deadlocks and livelocks. The switching mechanism determines how packets are forwarded and buffered in routers; it is a relatively simple method for NoCs that requires small buffers and has low latency. Flow control reserves network resources to prevent congestion and deadlocks via virtual channels, credit-based flow control and admission control. Together, the three will ensure high-efficiency and stable on-chip data transmission. Requests from computing nodes are divided into smaller segments by the network interface in practice, and then the source router selects an output port based on the routing algorithm. They pass through several routers with buffering, arbitration and crossbar switching before reaching the destination node, and then the network interface reassembles the data and delivers it.

3. Current research status of NoC technology

3.1. Topological structure

The topology determines the physical connection form and scalability limits of a NoC and serves as the foundation for on-chip network design. As the number of cores grows, researchers have proposed Torus, Ring, Fat-Tree, and other variant topologies to optimize communication performance in specific scenarios, mainly by shortening network diameter. After 3D stacking technology matured, 3D Mesh topology became a research hotspot, significantly reducing interconnect lengths and improving bandwidth density. In recent years, looking at the highly regular and localized data flow characteristics in neural networks, researchers have proposed specialized topologies to boost communication efficiency for AI workloads by reducing long-distance communication and matching data reuse patterns.

At the same time, the rise of Chiplet architectures has extended topology design from a single chip to the package level, using high-bandwidth parallel links or serial SerDes interfaces between

DOI: 10.54254/2755-2721/2026.CH37633

chiplets to achieve large-scale cross-chiplet expansion. These communication-centric interconnect fabrics are characterized by different trade-offs with regard to latency, throughput, energy dissipation, and silicon area requirements [9].

3.2. Routing algorithm

Routing algorithms are responsible for deciding the path that data packets take through the network, which directly affects the network's latency and load balancing. Deterministic routing, represented by XY routing, determines paths solely based on the coordinates of the source and destination nodes, and it's still the most commonly used routing scheme in commercial NoCs. However, XY routing can lead to localized congestion under uneven traffic distributions, so researchers have proposed a series of adaptive routing algorithms like DyAD and O1TURN, which significantly improve network throughput and load balancing under non-uniform traffic. LAXY is based on the fact that dimension-ordered routing algorithms imposes the highest traffic load on the central nodes in mesh topologies [10]. In Chiplet scenarios, the design space for routing algorithms expands even further. However, it is not easy to build interconnection networks across multiple chiplets and maintain high-performance deadlock-free routing in systems of various hierarchical topologies [11]. Figuring out how to achieve global congestion awareness and optimal path decisions across layers is still an open research question. Additionally, smart routing algorithms are starting to get attention—they train agents to dynamically decide routing strategies based on real-time network conditions and show better performance than traditional algorithms under complex heterogeneous traffic. Unfortunately, with this method, the implementation overhead of RL increases rapidly as the network scale increases [12].

3.3. Exchange mechanism

The switching mechanism defines how packets are forwarded and buffered inside a router. Wormhole switching is currently the most widely used switching mechanism in NoCs. Its buffer requirements are only related to the size of flits rather than the whole packet, which greatly reduces the buffer resource overhead of the router while keeping transmission latency low. The main problem of this traditional wormhole switching method is a head-of-line blocking problem. If the header flit is blocked then it will block the remaining paths that are still used by the wormhole packet [13]. To address this issue, researchers have proposed hybrid switching mechanisms combining virtual cut-through and wormhole switching, as well as customized switching solutions for AI data flows. For example, for scenarios where neural networks involve lots of small, regular packets, mechanisms supporting flit-level out-of-order forwarding and batch scheduling have been designed to improve switching efficiency under AI workloads. But in Chiplet architectures, the latency and power consumption of inter-chiplet links are much higher than on-chip links, and switching between coarse-grained inter-chiplet and fine-grained on-chip wormhole introduces more complexity in designing hierarchical switching mechanisms.

3.4. Flow control mechanism

The flow control mechanism is responsible for managing buffers and link resource allocation in the network, and it's a core system to ensure stable operation of NoCs. Virtual channel technology forms the basis of flow control. Each virtual channel has its own buffer and status, and they share physical link bandwidth through time-division multiplexing, which not only improves link utilization but also

DOI: 10.54254/2755-2721/2026.CH37633

fundamentally prevents deadlocks by breaking cyclic dependencies. That's why almost all modern NoC designs include virtual channels. In AI scenarios, flow control optimization focuses on matching the communication patterns of neural networks. For example, for AI inference data flows, a dynamic virtual channel allocation method has been proposed: during communication peaks, more virtual channels are assigned to AI traffic, and during idle times, resources are reclaimed to reduce power consumption. Our evaluation shows that our system reduces the communication delay by 55.47% on average, and reduces energy consumption of the system by 21% on average [14]. So the flow control mechanism improves resource utilization and energy efficiency under AI workloads. In Chiplet architectures, flow control needs to manage buffer resources both within chips and between chiplets. Researchers have proposed mechanisms like credit-based flow control for long-latency links and explicit congestion notification to improve inter-chiplet link bandwidth usage. New deadlocks may occur across both the chiplets and the interposer due to the integration, even if the NoC of each individually designed chiplet is deadlock free [15].

4. Limitations and prospects

4.1. Performance bottleneck in inter-core connections and lack of hierarchical collaborative design

The three types of deficiencies are as follows. First, the bandwidth density and latency of the interconnections between chiplets are significantly lower than those in-chip, and thus cross-chiplet communication becomes a performance bottleneck. Second, most current research treats the on-chip NoC and inter-chiplet interconnects as two separate layers for optimisation and lacks a unified end-to-end traffic scheduling mechanism; thus, it is difficult to achieve optimal data transfer across the entire path from compute cores to other chiplets. Third, the heterogeneous integration of multiple chiplets in a Chiplet architecture cannot meet the communication requirements of all chiplets simultaneously, and thus the overall system communication efficiency is reduced.

Therefore, the three directions for future research are as follows. One is studying optoelectronic hybrid interconnect technology to add optical links between chiplets and fundamentally increase cross-chiplet bandwidth and energy efficiency. The two are building a hierarchical, collaborative, integrated network architecture to jointly design the on-chip NoC and the package-level network (NoP) for global congestion awareness and end-to-end path optimisation. Three is developing configurable network architectures for heterogeneous chiplets and, using techniques such as parameterised virtual channel dynamic allocation, improving the overall resource utilisation and communication efficiency of multi-chiplet heterogeneous systems.

4.2. Insufficient dynamic load adaptability and lagging industrial ecosystem standardization

The deficiencies in the second area are also shown at two levels. First, AI reasoning and training workloads are highly dynamic, and most current NoCs cannot adjust network behaviour according to real-time load. Therefore, there is a considerable increase in communication delay during peak hours and a waste of network resources in off-peak hours. Second, the Chiplet industry ecosystem is still immature. Universal chiplet interconnect standards (such as UCIe) are still being developed, and differences in interface protocols, data formats and quality-of-service semantics among chiplets from different vendors are hindering the large-scale application of Chiplet architecture and the construction of a collaborative industry ecosystem.

DOI: 10.54254/2755-2721/2026.CH37633

To address the above deficiencies, future studies can take two paths. First, develop machine-learning-based intelligent adaptive NoCs that dynamically optimise routing strategies, virtual channel allocation and bandwidth scheduling according to real-time AI load characteristics to achieve adaptive and self-optimising network behaviour. Second, actively participate in the optimisation of universal chiplet interconnect standards such as UCIe, explore standardised layered interfaces and unified quality-of-service semantics for NoCs, and develop open-source Chiplet NoC design and validation platforms to accelerate the industrial application of Chiplet architecture and build an open ecosystem.

5. Conclusion

This paper presents an organised review and analysis of on-chip network (NoC) technology for AI chips in the context of chiplet integration. First, it introduces the general structure and hierarchical working mode of on-chip networks. Then, based on the four categories of topology, routing algorithms, switching mechanisms and flow control mechanisms, it systematically introduces the development of NoC research and customised applications for AI and Chiplet scenarios. The two main deficiencies in the paper are performance bottlenecks for inter-chiplet connections and a lack of hierarchical collaborative design; additionally, there is insufficient dynamic load adaptability and lagging industrial standardisation. In the future, this area will continue to develop along the path of optoelectronic hybrid interconnects, hierarchical collaborative architectures, intelligent adaptive networks and standardised industrial ecosystems. The research in this paper provides a full-featured technical reference and direction for Chiplet-based AI chip interconnect design, and serves as an important reference for improving the performance and industrial application of high-integration AI chips in the post-Moore era.

References

[1] Mutlu, O., et al. (2019). Processing data where it makes sense: Enabling in-memory computation. *Microprocessors and Microsystems, 67*, 28–41.
[2] Mahmud, M. T., & Wang, K. (2024). A flexible hybrid interconnection design for high-performance and energy-efficient chiplet-based systems. *IEEE Computer Architecture Letters, 23*(2), 215–218.
[3] Benini, L., & De Micheli, G. (2005). Network-on-chip architectures and design methods. *IEE Proceedings - Computers and Digital Techniques, 152*(2).
[4] Abadal, S., et al. (2013). Area and laser power scalability analysis in photonic networks-on-chip. *17th Conference on Optical Network Design and Modelling*, 131–136.
[5] Benini, L., & De Micheli, G. (2002). Networks on chips: A new SoC paradigm. *Computer, 35*, 70–78.
[6] Kumar, A., Murali, S., & Kamakoti, V. (2012). Network-on-chips on 3-D ICs: Past, present, and future. *IETE Technical Review, 29*(4), 318–335.
[7] Hao, Y., et al. (2025). Survey of chiplet technology: SoC architecture, interconnect, EDA, and advanced packaging. *IEEE Journal on Emerging and Selected Topics in Circuits and Systems, 15*(4), 514–536.
[8] Kumar, S., et al. (2002). A network on chip architecture and design methodology. *Proceedings IEEE Computer Society Annual Symposium on VLSI. New Paradigms for VLSI Systems Design. ISVLSI 2002*, 117–124. Pittaburgh, PA, USA.
[9] Pande, P. P., et al. (2005). Performance evaluation and design trade-offs for network-on-chip interconnect architectures. *IEEE Transactions on Computers, 54*(8), 1025–1040.
[10] Rohbani, N., et al. (2017). LAXY: A location-based aging-resilient Xy-Yx routing algorithm for network on chip. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, 36*(10), 1725–1738.
[11] Feng, Y., Xiang, Y., & Ma, J. (2023). A scalable methodology for designing efficient interconnection network of chiplets. *2023 IEEE International Symposium on High-Performance Computer Architecture (HPCA)*, 1059–1071. Montreal, QC, Canada.

DOI: 10.54254/2755-2721/2026.CH37633

[12] Wang, Z., et al. (2023). LARE: A linear approximate reinforcement learning based adaptive routing for network-on-chips. *2023 IEEE International Symposium on Circuits and Systems (ISCAS)*, 1–5. Monterey, CA, USA.
[13] Samman, F. A., Hollstein, T., & Glesner, M. (2011). Wormhole cut-through switching: Flit-level messages interleaving for virtual-channelless network-on-chip. *Microprocessors and Microsystems, 35*(3), 343–358.
[14] Fang, J., Chang, Y., & Li, X. (2020). Exploration on routing configuration of HNoC with intelligent on-chip resource management. *IEEE Access, 8*, 12117–12129.
[15] Wu, Z., et al. (2022). Upward packet popup for deadlock freedom in modular chiplet-based systems. *2022 IEEE International Symposium on High-Performance Computer Architecture (HPCA)*, 986–1000. Seoul, Korea.