Transactions on Computer Science and Intelligent Systems Research
ISSN: 2960-1800, eISSN: 2960-2238 | Volume 14
TPCEE 2026

Overview of Multi Chip and Heterogeneous Computing Architecture for AI Large Models

Jiawei He [1, *]

[1] Guangzhou SCA School, Guangzhou, China

* Corresponding Author Email: 2277402962@qq.com

Abstract. With the parameter scale of large-scale AI models breaking through the trillion level, traditional single-chip architectures are facing fundamental bottlenecks in terms of computing power expansion, storage bandwidth, and manufacturing yield. Multi-core and heterogeneous computing architectures integrate multiple functional chips into a unified system through advanced packaging technology, which has become a key technological path supporting the continuous growth of large model computing power in the post Moore era. This article systematically reviews the core technology system of multi-core heterogeneous computing architecture and summarizes the research status and technical challenges from four dimensions: 2.5D/3D advanced packaging technology, inter chip interconnect protocol, heterogeneous computing and in memory computing fusion architecture, and workload mapping methods. On this basis, the trend of evolution from "encapsulation integration" to "architecture interconnect software collaborative design" in this field was analyzed, and it was pointed out that core issues such as inter chip communication bottlenecks, cross chip workload partitioning, and software hardware collaborative optimization still need further research. This article aims to provide technical references for the design of specialized chips and systems for AI large models.

Keywords: Multi core particles; Heterogeneous computing; AI big model; Advanced packaging; Inter chip interconnection.

1. Introduction

The core idea based on Multi Chiplets is to decompose the traditional single-chip System on Chip (SoC) into multiple small chips with independent functions and different process nodes, and then integrate them into a complete system through advanced packaging technology. This technological route not only breaks through the limitation of lithography area, but also integrates heterogeneous chips such as digital computing chips, analog in memory computing chips, and high bandwidth storage chips (HBM) into the same package, achieving heterogeneous integration of "optimal process manufacturing and optimal functionality". For AI large model scenarios, multi-core architecture has become a common technology choice for mainstream AI accelerator products such as NVIDIA GB200 Grace Blackwell, AMD MI300X, Intel Gaudi-3, etc.

In terms of advanced packaging technology, TSMC's CoWoS (Chip on Wafer on Substrate) series technology has developed to the fifth generation, with an intermediate layer area of over 2500 mm ², which can integrate multiple logic chips and up to 8 HBMs; Intel's EMIB (Embedded Multi die Interconnect Bridge) technology achieves local high-density interconnection between chips through embedded silicon bridges, and the interconnection pitch has been reduced to 36 μ m; New technological routes such as glass based packaging are also being explored. At the level of inter chip interconnect protocol, UCIe (Universal Chiplet Interconnect Express) Standards are gradually becoming industry standard protocols, providing a foundation for interoperability between chips from different manufacturers. At the architectural level, FLARE et al. proposed a heterogeneous in memory computing multi-core granular accelerator architecture, which achieves LLM inference acceleration through fine-grained hardware software collaborative design; Stardust focuses on the workload mapping problem of multi-core systems, using deep reinforcement learning methods to optimize the cross core partitioning of computational graphs [2].

Content from this work may be used under the terms of CC BY-NC 4.0 licence (https://creativecommons.org/licenses/by-nc/4.0/).
Published by Warwick Evans Publishing.

However, the communication bottleneck between chips has not been fundamentally resolved. The bandwidth and energy efficiency of inter chip interconnection are much lower than those of on-chip interconnection, and cross chip communication caused by mechanisms such as Mixture of Experts (MoE) in large model inference has characteristics of suddenness and spatial imbalance, making it difficult to effectively cope with existing static interconnection resource allocation.

The computational scale of large models is enormous, and the heterogeneity between chips leads to an explosive growth in the search space for optimal partitioning problems. Existing heuristic or reinforcement learning methods still have limitations in scalability and transferability.

This article aims to systematically review the key technologies and research frontiers in this field, revealing the inherent logic and future directions of technological evolution. In the rapidly developing stage of multi-core heterogeneous computing, a clear knowledge framework is established for this field through systematic literature review and technical classification. At the same time, it provides technical roadmap references for architects of AI large model dedicated chips and systems.

2. Advanced packaging and integration technology with 2 multi-core chips

2.1. 2.5D Integration Technology

2.5D integration is the most mature packaging solution in multi-core systems. Its core feature is the introduction of an intermediate layer between the chip and the substrate, which achieves high-density interconnection between chips through metal wiring on the intermediate layer.

TSMC's CoWoS series technology is currently the most widely used 2.5D packaging solution. CoWoS-L technology utilizes local silicon interconnect chips and reconstruction intermediaries, ensuring high bandwidth and low latency performance while achieving large integration areas. NVIDIA's GB200 Grace Blackwell chip is based on CoWoS-L technology, integrating two Blackwell GPUs and one Grace CPU in the same package, supporting LLM with over 10 trillion parameters, equipped with 384GB of off chip memory, and a total device power consumption of 2700W [1].

AMD's MI300X showcased the technology roadmap of 2.5D and 3D hybrid integration. In this scheme, CPU chips (CCD) and accelerator chips (XCD) are placed on top of I/O chips through 3D stacking to achieve low signal delay, and then the entire 3D stack is integrated with HBM chips through a large-sized silicon intermediate layer.

Intel's EMIB technology adopts different paths. It does not use the entire intermediate layer, but embeds small silicon bridges in the packaging substrate, providing local interconnection only between the edges of the chips that require high-density interconnection. This method maintains sufficient interconnect density while reducing manufacturing costs. The Gaudi-3 accelerator integrates two computational chips and 128GB HBM using EMIB technology.

Glass based packaging is becoming a new direction for 2.5D integration. Compared with organic substrates and silicon interlayers, glass has a smoother surface, lower coefficient of thermal expansion, and better dimensional stability, which can support finer line widths and spacing and higher density interconnects. Intel has showcased its first glass substrate test chip, and Absoluts has also begun preparing small-scale production of glass substrates for ultra large scale customers.

2.2. 3D Stacking Technology

3D stacking connects multiple chips vertically through silicon via technology, achieving higher interconnect density and shorter signal paths than 2.5D. The vertical stacking of CPU and I/O chips in AMD MI300X is a representative case.

The main advantage of 3D stacking is that it significantly shortens the physical distance between chips, thereby reducing signal delay and power consumption. However, 3D stacking also faces challenges such as heat dissipation difficulties, high manufacturing costs of through silicon vias, and

thermal mechanical stress. Currently, 3D stacking is commonly used in conjunction with 2.5D intermediate layer technology: vertical connections between chips handle high bandwidth, low latency tightly coupled communication, while intermediate layers are responsible for global interconnection and peripheral connections [3].

2.3. Development Trends of Packaging Technology

The current packaging technology is evolving towards higher integration, lower cost, and better signal integrity. The CoWoS-R technology adopts an organic intermediate layer scheme, which connects the UCIe interface through a redistribution layer. Research has shown that when implementing UCIe-A IP on the CoWoS-R platform, special attention needs to be paid to the design of the power distribution network - due to the lack of embedded deep trench capacitors, external integrated passive device capacitors are required to suppress voltage fluctuations, which can suppress voltage fluctuations at the μ bump from 75.90mV to 21.98mV.

The exploration of emerging packaging technologies also includes integrating passive devices more tightly with chips, developing more efficient heat dissipation solutions, and exploring new substrate materials such as glass. These developments collectively point to a trend: encapsulation is no longer just a simple "connection" function, but has become a key link in optimizing system performance.

3. 3-core inter grain interconnect protocol and communication architecture

3.1. Mainstream D2D interface solutions

The inter chip interconnect interface is the core element that determines the performance of multi-core systems. At present, there are mainly two types of solutions in the industrial sector: proprietary interfaces and standardized interfaces.

AMD's Infinity Fabric is a proprietary D2D interface used in its multi-core products, designed to minimize latency and maximize bandwidth. Intel's AIB is also a proprietary interface for similar purposes. These proprietary interfaces perform excellently in specific products, but the lack of interoperability between chips from different manufacturers limits the flexibility of the ecosystem.

The emergence of UCIe standards has changed this situation. The UCIe 1.0 specification was released in 2022, defining the physical layer, protocol layer, and software stack for inter chip interconnection. A key feature of UCIe is its support for multiple packaging technologies, including standard packaging, organic interlayers, and silicon interlayers. The parasitic parameters of interconnect channels vary significantly under different packaging methods, so the specifications explicitly limit the voltage transfer function loss and crosstalk indicators.

3.2. UCIe Standards and Their Technical Challenges

UCIe faces technical challenges in both signal integrity and power integrity in practical engineering implementation. In terms of signal integrity, high-speed interconnect channels are susceptible to data corruption, loss, and distortion. The fine pitch interconnection in advanced packaging makes the signal line spacing smaller, the coupling effect between layers and chips stronger, and the crosstalk problem more prominent.

Regarding this issue, research has shown that using a ground signal ground staggered transmission line structure on the CoWoS-R organic intermediate layer can achieve low insertion loss (3.79) dB@18GHz) And low crosstalk (-28.24) dB@18GHz) At a speed of 36 GT/s, the eye diagram width reached 0.74 UI. This scheme validates the feasibility of UCIe implementing 64 bit, 16-36 GT/s high-speed interconnection on organic intermediate layer platforms [4].

In terms of power integrity, the design of the power distribution network is equally critical. Multiple chips in a chip system share a power distribution network, and the switching noise of one chip may

couple to another chip through the power network. The CoWoS-R platform lacks embedded deep trench capacitors and requires external integrated passive devices to provide local decoupling paths.

3.3. Evolution Trend of Interconnection Protocol

The evolution of UCIe 2.0 further enhances bandwidth capability. When implementing UCIe 2.0 64 bit routing on thin organic redistribution layers, wiring can be completed within a six layer stack, which reduces manufacturing costs while ensuring signal and power integrity compared to traditional 8-10 layer schemes [5].

From a broader perspective, standardized D2D protocols are crucial for design flexibility and scalability, especially in large AI and HPC systems. As more manufacturers support the UCIe standard, specialized chips from different manufacturers will be seamlessly integrated, greatly enriching the design space of multi-core systems.

4. Heterogeneous Computing and In Memory Computing Fusion Architecture

4.1. Basic ideas for heterogeneous chip integration

The computing modes of AI big models have a high degree of diversity: computationally intensive tasks such as matrix multiplication are suitable for running on GPUs or dedicated accelerators, control logic and scheduling tasks are suitable for execution on CPUs, and memory access intensive tasks benefit from the near data processing capabilities of high bandwidth storage. One major advantage of multi-core architecture is the ability to integrate different types of computing and storage chips into the same package, matching appropriate computing resources based on task types.

The NVIDIA GB200 integrates Grace CPU with Blackwell GPU, allowing the system to dynamically allocate computing tasks based on workload characteristics. The AMD MI300X achieves efficient support for large-scale parallel computing tasks by 3D stacking multiple GPU chips and I/O chips, and integrating them with HBM through an intermediate layer [6].

4.2. In Memory Calculation and Near Memory Calculation

In memory computing is an important technological path to solve the "memory wall" problem of AI large models. In traditional computing architectures, frequent data transfer between processors and memory has become the main bottleneck for performance and energy efficiency. In memory computing reduces the need for data handling by directly performing calculation operations in storage units.

In a multi-core architecture, in memory computing chips can be integrated as independent chips with digital computing chips. This heterogeneous integration solution needs to address interface issues between digital chips and analog in memory computing chips, process compatibility issues, and data accuracy matching issues. Although this direction is still in the early stages of exploration, it is considered an important technical direction for large-scale model inference.

Tesla's Dojo chip demonstrates another approach. This chip is specifically optimized for training large neural networks and achieves high-density, low latency system integration through TSMC's InFo SoW technology. The total device power consumption is only 400W, far lower than similar competitors. This system level wafer integration solution explores the path of computing storage fusion at a more macro level.

5. Workload partitioning and task mapping for large models

In multi-core granular systems, workload partitioning and mapping have a decisive impact on system performance. Due to the fact that the storage capacity of each chip is usually smaller than that of a single-chip accelerator, how to effectively partition and allocate large-scale tensor computation graphs to each chip directly affects the efficiency of model training and inference.

The main challenges faced by workload mapping for large models come from two aspects. One reason is the increase in the number of core particles, the enhancement of core particle heterogeneity, and the expansion of computational graph scale, which collectively lead to the problem of combinatorial explosion. Secondly, different models have different computational graph structures and communication modes, and the mapping scheme optimized on one model is often difficult to transfer to another model.

5.1. Method Analysis

In recent years, learning based methods have shown better potential. The Stardust framework combines learnable graph clustering, topological mask attention, and deep reinforcement learning: first, graph clustering is used to shrink large-scale computational graphs to a manageable scale, then topological mask attention is used to capture the structural information of computational graphs, and finally, a kernel mapping scheme is generated through deep reinforcement learning strategies. The evaluation on industrial scale AI models shows that the mapping scheme generated by Stardust significantly outperforms commonly used heuristic methods in terms of throughput, and fine-tuning the pre trained model can increase sample efficiency by 15 times from zero training [7].

Other studies are also exploring similar directions. CoSA et al. proposed a method for scheduling spatial accelerators through constraint optimization. Gemini focuses on the collaborative exploration of mapping and architecture. A research team from the University of California proposed a two-layer optimization method for multi model AI workload scheduling on heterogeneous chip modules, which includes two levels: time partitioning for chip reconfiguration and spatial mapping from sub models to chips [8].

5.2. Trends in Collaborative Optimization of Software and Hardware

The design of multi-core systems for AI big models is evolving from a serial mode of "designing hardware first, then mapping software" to a mode of "software hardware collaborative design". This means that workload mapping requirements need to be considered during the architecture design phase, and mapping algorithms also need to fully utilize the flexibility provided by hardware.

In a multi-core system, the packaging structure determines the physical interconnection topology between chips, and the interconnection topology directly affects the communication efficiency of workloads between chips. There is a close coupling relationship between the three. Previous studies have shown that the selection of packaging architecture for chip level systems not only affects signal integrity and power consumption performance, but also determines the maximum frequency and effective bandwidth limit for chip to chip communication. Therefore, packaging solutions and architecture planning must be considered together in the early stages of design [9] In addition, traditional high-density interconnect designs often rely on engineers' experience and extensive electromagnetic simulation iterations, with long design cycles and difficulty in covering a large parameter space. To address this issue, researchers have developed an optimization framework based on deep reinforcement learning, which can automatically search for the optimal geometric parameters and layout combinations of signal and ground lines in the CoWoS packaging process. While meeting the UCIe eye diagram and voltage transfer function specifications, it effectively suppresses crosstalk in high-speed data transmission and improves signal quality. This type of method shortens the design cycle from weeks to hours, demonstrating significant engineering application prospects [10].

6. Challenges and Prospects

6.1. Core Challenges Currently Faced

Based on the above analysis, the multi-core granular heterogeneous computing architecture for AI big models currently faces several core challenges. These challenges are constrained from three

dimensions: physical communication layer, system design layer, and engineering economy layer, and there is a deep causal relationship between them.

Firstly, the fundamental constraint at the physical level - inter chip communication remains the main performance bottleneck. Although standardized protocols such as UCIe and advanced packaging technologies continue to promote the improvement of interconnect bandwidth and energy efficiency, chip to chip communication relies on metal interconnects and intermediate layers in the packaging, and there is an order of magnitude gap between its signal transmission path, parasitic effects, and power consumption overhead and on-chip communication, which is physically difficult to eliminate. What is even more tricky is that the cross chip communication caused by mechanisms such as mixed experts in large models has characteristics of suddenness and spatial imbalance, and the currently statically configured interconnect resources cannot make dynamic adjustments according to real-time changes in communication traffic, resulting in frequent problems of interconnect bandwidth overload or idle in scenarios such as MoE inference, which seriously restricts the actual throughput of the system.

Due to the increasing complexity of cross level collaborative design. Encapsulation technology determines the physical characteristics of communication between chips, interconnect architecture specifies the implementation of communication protocols, chip layout affects the length and delay of communication paths, and workload mapping determines the distribution pattern of communication traffic in time and space. The coupling relationship between these four levels makes the entire design space extremely large, and any single level of local optimization may fail due to mismatches with other levels, making it difficult to achieve global optimality. However, if joint search is to be performed on four levels simultaneously, the design space increases exponentially with the number of chips and interconnect ports, resulting in significant computational overhead.

6.2. Future Development Trends

From the perspective of technological evolution, packaging technology will continue to evolve towards higher density and lower cost. Emerging solutions such as glass based packaging and organic interlayer optimization are expected to gradually mature and enter mass production. Organic intermediate layer solutions such as CoWoS-R have validated the feasibility of UCIe's 64 bit interface at 36 GT/s in terms of signal integrity, providing technical validation for future packaging solutions with higher bandwidth and lower cost.

In addition, AI driven design optimization will run through the entire chip design process. From chip layout optimization to workload mapping, from thermal management to power integrity analysis, machine learning methods are becoming the core tools for exploring high-dimensional design spaces.

7. Conclusion

This article systematically reviews the key technologies of multi-core and heterogeneous computing architecture for AI big models, and summarizes the research status and technical challenges from four dimensions: advanced packaging, inter chip interconnect protocols, heterogeneous fusion architecture, and workload mapping. Overall, multi-core heterogeneous computing architecture has become a key technological path supporting the continuous growth of AI large model computing power. With the collaborative advancement of encapsulation technology, interconnect protocols, and software methods, this field is expected to continue unleashing computing power growth space in the post Moore era.

References

[1] Heterogeneous Integration Technologies for Artificial Intelligence Applications. IEEE Xplore, 2024.

[2] Advanced semiconductor packaging design via artificial intelligence and machine learning: A review. ScienceDirect, 2026.

[3] Thermal Management Challenges in 2.5D and 3D Chiplet Integration: A Review on Architecture–Cooling Co-Design[J]. Applied Sciences, 2025, 6(12): 373.

[4] WANG W C, HUNG C C, LIN Y T, et al. Design of UCIe-A x64 Chiplets Integration for 16-36 GT/s Using Organic Interposer[C]//Proceedings of IEEE 76th Electronic Components and Technology Conference (ECTC). IEEE, 2026.

[5] KUMAR A, SHAH J, R N. High-bandwidth UCIe 2.0 routing on fine-line organic RDL with enhanced signal and power integrity for scalable 64-bit multi-chiplet integration[C]//Proceedings of 2026 IEEE 76th Electronic Components and Technology Conference (ECTC). IEEE, 2026: 1826-1831.

[6] NVIDIA. The NVIDIA Grace Blackwell Superchip[EB/OL]. (2025-04-20)[2026-08-24].

[7] LI T, HOU J, YAN J L. Stardust: scalable and transferable workload mapping for large AI on multi-chiplet systems[J]. IEEE Computer Architecture Letters, 2025, 24(2): 201-204. DOI: 10.1109/LCA.2025.3580562.

[8] HUANG Q J, KANG M W, DINH G, et al. CoSA: scheduling by constrained optimization for spatial accelerators[C]//Proceedings of the 48th Annual International Symposium on Computer Architecture (ISCA). IEEE, 2021.

[9] Lu Meixuan, Xu Haobo, Wang Ying, etc Chip Integrated Chip Architecture Packaging Collaborative Design [J]. Journal of Electronics and Information Technology, 2025, 47 (9): 2994-3009

[10] MIAO W Y, TAN C S. Signal integrity optimization for CoWoS chiplet interconnection design assisted by reinforcement learning[C]//Proceedings of 2024 IEEE 10th Electronics System-Integration Technology Conference (ESTC). IEEE, 2024: 1-6