CMOS-Based AI Chip Design and Development Trends

Letao Gong

College of Engineering, Northeastern University, Boston, MA, 02115, United States

gong.let@northeastern.edu

Abstract. With the rapid growth and development of end devices, CMOS-based chips have encountered a bottleneck in development. Traditional CMOS manufacturing methods are reaching the physical limits based on Moore's law. In particular, the demands of technologies such as Artificial Intelligence, 5G and autonomous driving have driven the continuous advancements of chip design in recent years. By using traditional methods, when CMOS size continues to shrink, power consumption and heat dissipation issues have become increasingly prominent, limiting further improvement and future development of chip performance. There is increasing demand for higher performance, including inference capabilities and deep learning, conflicts with lower consumption and maintaining low costs. Therefore, finding a balance between power consumption, heat dissipation, and performance has become a direction for research and development of chip design. Success will depend on cross collaboration between different levels in the industry that includes both innovation and manufacturing, how to break through the bottleneck becomes a challenge in developments. CMOS-based AI chips will be an indispensable part in the future of high-tech industries, thus supporting the long-term developments in various industries.

Keywords: CMOS, AI chips, CMOS architecture, Cloud-edge collaboration, Advanced packaging.

1. Introduction

Reducing the size of CMOS is no longer satisfying the development needs, innovation must take place in architecture, packaging and design methods to adapt to the fast development. Ther emergence of new fields has higher requirements for chips, for example, artificial intelligence requires large-scale computing and data processing, medical electronics emphasize miniaturization, consumer electronics, such as mobile phones and tablets, require chips to achieve a balance between high integration and innovation. CMOS-based AI chips are not only an industry trend, but chip design is also driven by AI itself, such as AI for EDA, to accelerate circuit layout and optimize energy consumption. This article will review the design and development trends of CMOS-based AI chips in terms of process, architecture, packaging, and applications, providing a reference for development and trends in the post-Moore era.

2. Background: CMOS and artificial intelligence chips

Traditional Artificial Intelligence chips are based on CMOS manufacturing, while CMOS technology is still the core manufacturing process of chips due to its advantages of low power consumption and integration. As the process develops, traditional CMOS are facing bottlenecks in power consumption and reliability, which require new optimization methods to solve the problem, enabling it to complete complex calculations at a lower power consumption [1]. Such issues have become challenges to continue reducing the size of CMOS and support AI tasks. Some methods are used to address such energy efficiency challenges, including reduce power consumption and enhancing reliability and optimizing architecture and circuits. CMOS will not be replaced by new emerging technologies but will develop in coordination with such technologies to form a hybrid method called "CMOS+new process". In the future, CMOS will become more platform technology rather than the only source of performance, the shift will profoundly impact the design of such CMOS-based AI chips.

3. Design trends in cmos-based ai chips

3.1. Heterogeneous computing architecture

a) Role in CPUs

CPUs are mainly used when there are sequential executions and high single-threaded performances, which are suitable for control tasks and complex logical operations, also responsible for task allocations, data flow management, and interactions with I/O and memory [2]. Although the advantages of CPU computing have gradually weakened as the development of AI, it still has an impact on cross-architecture coordination. As CMOS technologies approach their physical limits, especially the reduction in chip size based on Moore's law, future CPU design is more likely to focus on power control section of CMOS-based AI chips and scheduling efficiency instead of high performances.

b) Role in GPUs

GPUs are composed of multiple stream processing units, which are eligible to execute thousands of threads simultaneously. Hence, they can do matrix operations, vector processing, image-based, and scientific computing. Hierarchical storage is used to reduce delays during task execution. Usually, there are collaborations with CPUs, while the GPUs are responsible for massive parallel computing. In the future, more collaborations with take place between GPUs, NPUs, and FPGA, while NPUs will focus on deep learning tasks and FPGA will execute customized logic, forming a multi-level structure with clear divisions. As GPU power consumption becomes a major issue, increasing the number of GPUs is not sustainable. Future designs of CMOS-based chips will focus more on improving memory architecture, storage optimization, and more energy-efficient circuits

c) Specialized accelerators: NPU and FPGA

Neural Processing Unit (NPU) is designed for AI to optimize the neural network computations, particularly the matrix multiplication and inference for deep learning models. NPUs usually have higher energy efficiency and throughput than CPUs and GPUs. However, the downside is its lack of flexibility. When there are significant changes to algorithm structure, the NPUs need to update their hardware support. Therefore, the NPUs' application will be more concentrated in cloud AI services and mobile chips.

A Field-Programmable Gate Array (FPGA) is composed of logic blocks, programmable interconnects, and I/O cells, which enable customized hardware circuits for applications. But due to high development complexity, there are limitations. With the development of High-Level Synthesis tools, FPGA development will be easier and more widely used in edge computing section of CMOS-based AI chips.

3.2. Chiplet modular design

a) Power and cost advantages

Chips from different manufacturing methods are used in chiplet technology. Traditional system-on-chip (SoC) usually relies on monolithic crafts. All modules, regardless of their performance requirement, must be in the same advanced process; some functions, such as battery management, are not required to have high performance, are forced into the area, leading to overspecification and unnecessary manufacturing costs [3]. While chiplets reduce the overall manufacturing cost by putting different sections in different processes, such as placing high-performance units on 5nm processes while some less demanding modules to 28nm processes, which is an advantage against the traditional SoC chip and can be more cost-effective.

Large scale SoCs require more interconnection to link function modules that are distributed across the chip, to increase signal transmission distance and have higher dynamic power consumption. In contrast, chiplet-modularization can largely reduce the losses during signal transmissions and related dynamic power usage. This infrastructure not only reduces power consumption but also optimizes the data transmission path to improve overall performance.

Moreover, modules can be produced independently by different manufacturers, which allows specification in certain areas of supply chain and enhances production efficiency, to use the most suitable processes for each module, and ultimately improves cost-effectiveness.

b) Scalability and flexbility in system integration

Small-sized chips significantly improve the utilization rate of the wafer, more chips can be manufactured when compared with a large monolithic SoC. Moreover, the reduction of the single chip area lowers the probability of defective chips. Manufacturers can increase production efficiency more extensively by discarding defected chips individually without affecting the entire system. Manufacturing waste can be reduced based on such method, enhancing flexibility and maximizing resource efficiency. Due to the modular nature of chiplets, it can achieve customization and optimization based on different needs. Multiple small chips are reorganized and recombined to support the domain-specific customization feature. Verified chiplets can be reused in different system configurations, to enhance the system scalability. Chinese Academy of Sciences proposed a "Big chip" can be possibly expanded from 16 to 100 chips, which is based on an extendable tile-based architecture. Each chiplet will integrate 16 RISC-V instruction set-based CPU processors and be interconnected via a network-on-clip. The design of symmetrical interconnection will support the high-bandwidth and low-latency between chiplets, to solve the problem that occurs in traditional monolithic SoCs. The processor also applied a unified memory system, and access is granted from any tiles across the entire processor.

c) Advanced packaging technologies

2.5D and 3D packaging provide the possibility for heterogeneous integration and enable high-performance multi-chip interconnection and system efficiency. 2D packaging is suitable for traditional multi-chip modules, and based on 2D packaging, 2.5D packaging achieves higher interconnect density, while 3D packaging provides shorter interconnect path, higher bandwidth density, and greater energy efficiency [4]. Chips of different processes and different function types include 2.5D and 3D technologies, are usually in the same system and can be efficiently integrated into the same package. The advanced packaging technologies exceed the limitations of monolithic SoCs. Also, the choice of interconnection is closely related to packaging technology. In 2.5D and 3D architecture, thermal design and power distribution should be considered.

3.3. Cloud and edge computing collaboration

a) Low latency and real-time decision-making

When doing edge computing, it is close to the data source, which can significantly reduce data transmission latency, and enable real-time reasoning and decision-making [5]. The high latency of traditional cloud computing architecture makes it difficult to meet the requirements of real-time applications such as autonomous driving and industrial control in CMOS AI applications. When comparing and quantifying the latency differences between cloud computing (100-500ms) and edge computing (1-20ms), latency has been reduced from seconds to milliseconds in smart manufacturing, and from minute to second in medical testing area when AI chips are used [6].

For AI chips, low latency is important to support real-time decision-making. Also, one of the core features of edge computing is that it reduces network transmission and shortens data paths. Low latency not only improves efficiency but also fundamentally guarantees user experience and security. In the future, latency can be further reduced by combining with AI accelerators.

b) Power efficiency in edge devices

Edge devices are limited by battery capacities and heat dissipation capabilities and require light and lower power AI models. Cloud devices can do the training tasks, and the edge devices will only perform the inference, which can reduce energy consumption. Some proposed technologies include model compression and quantization to enable edge devices to execute AI tasks within power constraints. Future optimization should not only rely on model-level technologies but also should explore the circuit level design, such as the low-power CMOS architecture and integrate some specialized accelerators such as NPUs, which can further improve power efficiency. In edge devices,

design algorithms with lower power consumption characteristics or dynamic voltage and frequency scaling (DVFS) to reduce energy consumption. When deploying AI on edge devices, it tends to quickly deplete batteries, power efficiency optimization is required to extend the operating time of the device [7]. Future research directions should more focus on circuit and architecture levels, for example, low-power CMOS architecture, reducing energy consumption by optimizing transistor size, and considering low power characteristics in the algorithm design stage, to achieve low power consumption through overall optimization of hardware and software, to achieve a balance between power efficiency and real-time performance.

c) Privacy and data security advantages

Edge computing will process data locally, reducing the need to upload sensitive information to the cloud and thus reducing the risk of data leakage. Some common technologies include encryption and access control. Edge computing keeps data storage and processing local to the device, allowing for more direct control and maintaining security policies. For Multi-Access edge computing (MEC), data is distributed across different nodes, which is less vulnerable to large-scale attacks and centralized cloud servers. Then security mechanisms are embedded during the system design phase, therefore it offers greater privacy protection advantages, mechanisms include Attribute-Based Access Control (ABAC) and Role-Based Access Control (RBAC), ensuring greater security of data and protecting user identities in a distributed environment [8]. Strong encryption mechanisms are critical to ensure secure communication and stored data in edge environments. But distributed edge devices are vulnerable to physical attacks then require security protocols and access controls to protect privacy and security. Edge computing can process sensitive data locally, but cloud processing is also indispensable, such as for large-scale data storage. Thus, effective privacy protection requires edge computing and cloud computing to work together. In the mechanism, some encrypted access control will increase the cost, so future research should focus on efficient and lightweight algorithms.

4. Optimization directions

4.1. Low power optimization: DVFS and Multi-threshold CMOS

a) Principle of DVFS

Dynamic Voltage and Frequency Scaling (DVFS) is used for saving power and optimizing performance. It will balance real-time workload and performance by adjusting its operating voltage and frequency dynamically to achieve a balance between power consumption and performance. Power consumption is proportional to voltage and frequency, so lower voltage and frequency can reduce dynamic power consumption [9]. When the system needs to perform large computing tasks, DVFS will increase voltage and frequency to ensure speed and responsiveness. If the system has a light workload, DVFS will reduce the voltage and frequency to have lower energy consumption and heat generation. This mechanism ensures that the chop can meet performance requirements while minimizing power consumption.

b) Multi-threshold CMOS for leakage reduction

Multi-threshold CMOS mixes high-threshold voltage and low-threshold voltage transistors in the same circuit. While high-threshold transistors are used to reduce static power consumption due to low leakage current, low-threshold transistors are used in critical paths to ensure speed and performance. Multi-threshold CMOS can effectively reduce overall power consumption while maintaining high-performance, which can achieve a balance between them. In circuit design, the distribution of high-threshold and low-threshold is a noticeable problem; excessive use of high-threshold CMOS may lead to performance degradation, while excessive use of low-threshold CMOS may weaken the optimization effect. As the CMOS sizing continues to shrink, it leads to more serious leakage problems. Therefore, multi-threshold CMOS will be more widely used in the future, and it will also face more challenges.

c) DVFS in edge AI processors

Dynamic Voltage Frequency Scaling (DVFS) will reduce voltage and frequency at low loads to reduce dynamic power consumption and increase voltage at high loads to avoid delays and saturation. Some DVFS accelerator frameworks are proposed, including DVFS predictor, dynamic V/F generator, PE array, and other modules [10]. In GPU, DVFS is an important resource management tool for mobile devices; it is used to reduce battery consumption while maintaining a certain inference speed. DVFS adjustment can find the optimal point in the three-dimensional space of energy consumption, fps, and latency. This characteristic is most suitable for edge AI processors, which need real-time performance and low power consumption. DVFS is a key technology for edge AI processors, used to maintain real-time inference capabilities under energy constraints.

4.2. Thermal management and new material

a) Thermal challenges in CMOS AI chips

The stacking method of multiple chips results in severe overheating problems, and the uneven heat dissipation causes material fatigue and reduces packaging reliability. In CMOS AI chips, which utilize heterogeneous structures like CPUs, GPUs, and accelerators, the uneven power distribution complicates thermal design. For materials, it is difficult for thermal interface materials (TIMs) to further reduce their thermal resistance; even improved TIMs can only reduce thermal resistance by approximately 20%. Large variations in thermal conductivity within the package layer led to uneven heat diffusion [11]. As CMOS size continues to shrink, the power consumption of individual transistors has decreased, but the chip size and number of transistors have increased significantly also the overall computing power used for AI has increased. This has led to thermal issues becoming a bottleneck earlier than power consumption issues.

b) High-K dielectrics in reducing leakage power

As CMOS size shrinks, leakage power becomes increasingly problematic. New materials are required to suppress short channel effects (SCE) and reduce power consumption, such as using High-K dielectrics in dual-gate FinFETs can limit current flow through the gate, resulting in low leakage current, get higher driven current, and improve electrical performance. High-K dielectrics can provide improved electron mobility, making them suitable for efficient operation at low voltages, also they can reduce leakage current in short-channel devices and improve the Ion/Ioff ratio [12]. However, quality will become a challenge in the future, when High-K dielectric contact silicon channel, it will cause defects, these defects may offset some of the performance improvements originally brought by High-K, so how to optimize the process to improve quality is the focus of development. Moreover, some High-K dielectrics will change at high temperatures, resulting in performance degradation, and then stability and manufacturability will become a problem. Reliability is also worthy of attention; for example, the different applications of High-K materials in edge computing devices and performance computing require differentiated changes based on different terminal needs.

c) New material compatibility and manufacturability

For the optimized silicon-based materials, a low-cost production method is used to manufacture silicon thin films on quartz substrates by using aluminothermic reduction, which solves the problem of reducing manufacturing costs while maintaining high quality. Then a method is proposed for environmental performance prediction and optimization in metal processing, which emphasizes that future materials must consider energy consumption, emissions, and environmental loads [13]. Even with large-scale manufacturing capabilities, such as the application of chemical vapor deposition (CVD) and metal organic chemical vapor deposition (MOCVD) in large quantities of wafer synthesis with a yield rate of 99%, uniformity and defectivity are still challenges. How to scale up the industry will become future issue [14]. The focus of future development should be on finding low-defect synthesis methods, developing more controllable and environment-friendly manufacturing processes, and establishing standardized quality and reliability systems to ensure the application of actual products.

4.3. Advanced packaging and high-bandwidth memory

a) Fundamentals of advanced packaging technologies

The earliest foundation of advanced packaging is flip-chip bumping, including the traditional C4 bump and the finer pitch C2 bump. These techniques determine the interconnection method between the chip and the substrate, affecting electrical and thermal performance, as well as the packaging density. Then, Hybrid Bonding (DBI) technology is used, which combines oxide-oxide bonding with metal-metal bonding to achieve connections at extremely low temperatures, which is also one of the most important foundational technologies in the future, enabling higher-density and lower-power packaging. Then, 2D integration is introduced by placing multiple chips on a single substrate, and silicon bridges are added to the substrate for 2.1D to increase interconnect density. Higher interconnect density is achieved by using TSV interposers for 2.5D integration, then 3D integration, and TSV stacking is used to provide vertical integration [15]. As the development of integration, the essence is that the interconnection density continues to increase, and the cost and process complexity gradually increase. Although 2.5D and 3D can bring huge performance improvements, high cost and low yield are still the main problems for popularization. In the future, hybrid architecture may become the mainstream in advanced packaging, which uses both 2D/2.1D technology and 2.5D/3D integration solutions in the same product. This layered strategy can achieve a balance between performance, power consumption, and cost.

b) Bandwidth and power efficiency benefits

The high bandwidth enables simultaneous transmission and rapid processing of large amounts of data, making it suitable for AI and High-Performance Computing (HPC). In the future, high bandwidth memory will further increase bandwidth through wider I/O and high-speed SERDES IP, rather than solely rely on the increase of frequency. Such technology uses a logic process to reduce voltage by saving 50% energy, advanced packaging to improve heat dissipation, and the addition of power TSVs to enhance energy efficiency [16]. AI and HPC require frequent exchange of large data sets when processing training and inference tasks. Higher bandwidth can meet this requirement for real-time and massively parallel computing. Lower power consumption also reduces heat, contributing to improving stability and efficiency.

While High-Bandwidth Memory has significant advantages in bandwidth and power efficiency, its cost and complex packaging process will prevent it from widely spreading. The requirements of bandwidth and efficiency for AI inference and AI training are different; training is more focused on bandwidth, while inferences are more in-edge devices, which are more focused on efficiency and costs. In the future, more HBM can be used in cloud end, at the same time, edge AI chips will adopt cheaper and lower efficiency solutions. A tiered market can be possible in the future, large AI companies and high-end products will use HBM, and edge devices will be hybrid storage solutions. HBM will not replace traditional memory technologies; instead, it will become a complement to other storage technologies, and the design will be driven differently by different application scenarios. Achieving a balance between performance and cost will be critical in the future of AI chip design.

c) 2.5D vs 3D packaging approaches

Advantages of 2.5D packaging, including lower cost and mature process, which is already widely deployed, are more suitable for large areas and multi-parallel interconnections. For 3D packaging, they will have higher interconnect density and energy efficiency, such as Foveros and Hybrid Bonding Interconnect (HBI), which can support logic-logic, logic-memory, thus shortening interconnect distance and increasing bandwidth. A new solution called Co-EMIB combines the advantages of 2.5D and 3D to achieve a unified architecture of horizontal and vertical interconnection [17]. 2.5D packaging will remain mainstream in the short term because it achieves a better balance between performance, yield, and cost. And 3D stacking is considered a long-term development direction because it can achieve higher functional integration in small package areas [18]. For power and cost-sensitive applications that require high-bandwidth interconnects, 2.5D will remain a more economical option. For end devices and AI accelerators, 3D packaging is more attractive due to its compactness and high interconnect density. Such converged architectures, such as Co-EMIB, can be

the trend in the future, while ensuring the yield rate of 2.5D, adding the close interconnection advantages of 3D to achieve an interconnection platform in both horizontal and vertical directions. These advanced packaging approaches not only improve data transmission efficiency but also help manage power and thermal challenges by providing a better heat dissipation path.

5. Conclusion

In conclusion, the article reviews and summarizes the possible design trend and development of CMOS-based AI chips in the future. The review focused on six directions: heterogeneous computing architecture, chiplet modular design, collaboration between cloud and edge, low power optimization, dynamic voltage frequency scaling and multi-threshold CMOS, thermal management, use of new materials and advanced packaging technologies. CMOS will remain as the mainstream in the AI chips industry, for the industry, it needs to solve the problems such as manufacturing cost and complexity of system integration.

Overall, although CMOS scaling already reaches the physical limit and faces the bottleneck, its important position in AI applications will be maintained by continuous innovation, including optimization, use of new materials, and advanced packaging. CMOS AI chips will continue to evolve and remain competitive in the future, while more collaboration between different areas is needed.

References

[1] Hsu, Yu-Chin, and Robert Chen-Hao Chang. 2020. "Intelligent Chips and Technologies for AIoT Era." In Proceedings of the IEEE Asian Solid-State Circuits Conference (A-SSCC), November 9–11, 2020, Online. IEEE.

[2] Vaithianathan, Muthukumaran. 2025. "The Future of Heterogeneous Computing: Integrating CPUs, GPUs, and FPGAs for High-Performance Applications." International Journal of Emerging Trends in Computer Science and Information Technology 1 (1): 12–23.

[3] Gujar, Vivek. 2024. "Chiplet Technology: Revolutionizing Semiconductor Design – A Review." Saudi Journal of Engineering and Technology 9 (2): 69–74.

[4] Li, Shenggao, Mu-Shan Lin, Wei-Chih Chen, and Chien-Chun Tsai. 2024. "High-Bandwidth Chiplet Interconnects for Advanced Packaging Technologies in AI/ML Applications: Challenges and Solutions." IEEE Open Journal of Solid-State Circuits and Systems 4 (December): 351–364.

[5] He, Wen. 2023. "Analysis of CMOS IC-Based Hybrid Architecture for Edge Computing." International Journal of Communication Networks and Information Security 15 (4): 134–148.

[6] Brahmaji, Kanagarla Krishna Prasanth. 2024. "Edge Computing and Analytics for IoT Devices: Enhancing Real-Time Decision Making in Smart Environments." International Journal for Multidisciplinary Research (IJFMR) 6 (5): 1–9. IJFMR.

[7] Hossain, Md Emran, Md Tanvir Rahman Tarafder, Nisher Ahmed, Abdullah Al Noman, Md Imran Sarkar, and Zakir Hossain. 2022. "Integrating AI with Edge Computing and Cloud Services for Real-Time Data Processing and Decision Making." International Journal of Multidisciplinary Sciences and Arts 19 (10): 1–15.

[8] Liu, Siqin, and Avinash Karanth. 2021. "Dynamic Voltage and Frequency Scaling to Improve Energy-Efficiency of Hardware Accelerators." 2021 IEEE 28th International Conference on High Performance Computing, Data, and Analytics (HiPC), 232–241.

[9] Zhang, Yiqun, Shuai Zhang, and Luyao Feng. 2024. "Research on Low-Power Microprocessor Design and Optimization Technology." In Proceedings of the 2024 International Conference on Power, Electrical Engineering, Electronics and Control (PEEEC), 362–366. IEEE.

[10] Lim, Jeong-A., Joohyun Lee, Jeongho Kwak, and Yeongjin Kim. 2024. "Cutting-Edge Inference: Dynamic DNN Model Partitioning and Resource Scaling for Mobile AI." IEEE Transactions on Services Computing 17 (6): 3300–3315.

[11] Katari, Monish, Jawaharbabu Jeyaraman, Ikram Ahamed Mohamed, and Kumaran Thirunavukkarasu. 2023. "Addressing Power and Thermal Challenges in Advanced Packaging for AI CPUs/GPUs: Insights

into Multi-die Stacking Technology." International Journal for Multidisciplinary Research (IJFMR) 5 (6): 1–15. IJFMR.

[12] Kumar, Jeetendra, Shilpi Birla, and Garima Agarwal. 2023. "A Review on Effect of Various High-K Dielectric Materials on the Performance of FinFET Device." Materials Today: Proceedings 79: 297–302.

[13] Zhu, Enbo, Zhengwei Zhang, and Chen Wang. 2023. "Editorial: Emerging Chip Materials and Devices for Post Moore's Era." Frontiers in Materials 10: 1224537.

[14] Katiyar, Ajit Kumar, Jonggyu Choi, and Jong-Hyun Ahn. 2025. "Recent Advances in CMOS-Compatible Synthesis and Integration of 2D Materials." Nano Convergence 12 (11).

[15] Lau, John H. 2022. "Recent Advances and Trends in Advanced Packaging." IEEE Transactions on Components, Packaging and Manufacturing Technology 12 (2): 228–247. IEEE.

[16] Kim, Kwiwook, and Myeong-jae Park. 2024. "Present and Future Challenges of High Bandwidth Memory (HBM)." In Proceedings of the 2024 IEEE International Memory Workshop (IMW), 1–4. IEEE.

[17] Sheikh, Farhana, Ramune Nagisetty, Tanay Karnik, and David Kehlet. 2021. "2.5D and 3D Heterogeneous Integration: Emerging Applications." IEEE Solid-State Circuits Magazine 13 (4): 77–87.

[18] Razdan, Sandeep, Jie Xue, Peter De Dobbelaere, Aparna Prasad, and Vipul Patel. 2022. "Advanced 2.5D and 3D Packaging Technologies for Next Generation Silicon Photonics in High Performance Networking Applications." In Proceedings of the 2022 IEEE 72nd Electronic Components and Technology Conference (ECTC), 428–435. IEEE.