Proceedings of CONF-FMCE 2026 Symposium: Smart City and Infrastructure Engineering
DOI: 10.54254/2755-2721/2026.CH37489

The Functions of Architecture of System-on Chips and Development Prospects of Combining with Artificial Intelligence

Kangxi Feng

Wuhan Britain-China School, Wuhan, China
kangxifeng2025@163.com

Abstract. In the current era of the 21st century, as the world's population has already reached a historic milestone of 8 billion, the urban scale has expanded and the population density of cities has tremendously risen. Similar to the trend of urban construction, the chips have also developed. For instance, the SoC (System-on-Chip) has some key additions, including upgraded ray tracing-supportive GPUs, new CPU microarchitectures and dedicated artificial intelligence accelerators (NPUs) for on-device inference. Nowadays, specialised SoCs for autonomous driving have also emerged. From a thinly populated town to a densely packed metropolis, the engineers are not simply piling more components onto chips; rather, they wisely use the powerful capabilities of AI and combine them together. AI SoC solutions have found broad applications in a wide range of fields, from vehicular perception systems on cars and mobile voice interfaces on phones to unmanned aerial vehicle (UAV) vision-based navigation. Despite there being several obstacles for the further development of AI SoCs, such as the limited area of the chips, leveraging AI to optimize chip design, resource scheduling, and thermal management could literally achieve the ultimate "AI for SoC". This paper aims to forecast the future development of AI SoC convergence, and offer practical recommendations to address potential bottlenecks such as thermal management, security vulnerabilities, and design complexity.

Keywords: System-on-Chip, Artificial Intelligence Accelerator, On-device Inference, Chiplet

1. Introduction

The System-on-chips (SoC) design integrates various function blocks such as accelerator, processor, memory, power managment and various interfaces on a single silicon chip. This architectural shift has enabled the adoption of AI across all the SOC design domains, including autonomous driving, smartphone voice recognition, and drone vision systems. Although recent research has made remarkable progress to explore the potential of deep learning and generative AI in addressing SoC design challenges, there are still obvious gaps in software and hardware collaborative optimization, low-power design and end-side adaptability [1]. Therefore, further exploring the deep integration of AI and SoC is of great significance to unleashing the potential of future chips [2].

© 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.CH37489

This paper aims to investigate the current capabilities and future prospects of AI SoC integration, particularly focusing on addressing these underexplored research possibilities. By using the method of literature review, this study summarizes the application of SoC technology in smartphones, automotive sensors, and drone navigation. This paper aims to forecast the future development of AI SoC convergence and offer practical recommendations to address potential bottlenecks like thermal management, security vulnerabilities, and design complexity.

2. Overview of SoC

2.1. Basic concepts

An SoC is a combination of various smaller chips such as CPUs or GPUs.It integrates the main computing, control, communication, and interface functions required to accomplish the target task onto a single chip platform. Traditional electronic systems often use different chips responsible for separate functions—such as CPUs, graphics chips, storage controllers, and signal processing chips—which are connected through circuit boards. An SoC, on the other hand, aims to integrate these functions into internal functional modules or Intellectual Property (IP) cores of the chip, and then connects them uniformly through an on-chip bus or network. Therefore, a single SoC typically possesses the basic capabilities required to run an operating system, receive sensor data, execute algorithms and control external devices. Table 1 illustrates the main components and corresponding functions of the SoC.

An SoC and a CPU cannot be treated as equivalent entities. The CPU is a processor that executes general instructions, while an SoC is a complete computing system that incorporates a CPU. Similarly, an NPU is not equivalent to an AI SoC. An NPU is only responsible for matrix multiplication, convolution, or tensor operations in neural networks and requires the CPU to complete task scheduling, the storage subsystem needs to provide weight and feature data, and the compiler, driver and runtime need to map the model to the hardware. Only when these modules form a coordinated system can it be called an AI SoC [3].

Table 1. Main components and functions of the SoC

	Functions	Significance in the system
CPU	Run the operating system and control programs	High flexibility, responsible for processes, exceptions and task scheduling
GPU/NPU/DSP	Parallel computing and dedicated acceleration	Improve the efficiency of graphics, signal processing and Al computations
Storage subsystem	Cache, on-chip SRAM and external DRAM interface	Determine whether the models and data can be continuously supplied
NoC [4]	Functional modules on the connector chip	Affecting bandwidth, latency, congestion and energy consumption
Interface	Connecting the camera, the network, the display and the actuators	Enable SoC to be embedded in real devices and control systems
Safety and power management	Trustworthy startup, isolation, DVFS and clock control	Ensure reliability, battery life and data security

DOI: 10.54254/2755-2721/2026.CH37489

2.2. Importance of SoC

An SoC functions similarly to a folder in organizing disparate elements: as the number of components increases, it groups them together to prevent scattering and inefficiency. Similar to how standalone components function, if units such as CPUs or GPUs operate independently, the overall system efficiency may be significantly reduced. Therefore, an SoC integrates them to operate as a cohesive system. Firstly, the internal connections are shorter than those at the board level, which can reduce pins, packaging and communication delays. Secondly, different modules can share storage, clock, power management and security mechanisms, thereby reducing size and energy consumption. Thirdly, designers can configure functions around specific scenarios such as mobile phones, automobiles, robots, industrial control or wearable devices, forming a "general-purpose processor + dedicated accelerator" combination [5]. The trade-off is an increase in design complexity: any mismatch across modules, interconnections, or software stacks may prevent theoretical peak performance from translating into real-world application performance.

2.3. Applications of SoC in daily life

An System-on-Chip (SoC) has become the "brain" of modern electronic devices by highly integrating computing, graphics, and communication modules [6]. In mobile terminals, SoCs enable high-performance computing and 5G connectivity in smartphones and tablets [7]. In the Internet of Things and smart homes, they provide low-power processing and autonomous sensing capabilities for devices such as smart speakers and robotic vacuum cleaners. In the automotive sector, SoCs serve as the core computational engine for intelligent cockpit entertainment systems and autonomous driving assistance decision-making. Moreover, with the rise of AI PCs, SoC architectures have been adopted in laptops, enabling efficient local operation of large AI models through integrated NPUs [8]. At the same time, the high integration of SoCs ensures long battery life and diverse functionality in space-constrained wearable devices such as action cameras, drones, and smart rings [9]. As AI applications become increasingly widespread, integrating AI directly into SoCs enables devices to be smarter, respond faster, and consume less power, highlighting the necessity of combining AI with SoC.

3. The combinations of AI and SoC

3.1. AI on SoC

The first combination of SoC and Al is to deploy the AI model onto the SoC, which is referred to as "Al on SoC". Early solutions often relied on the CPU to call external GPUs or FPGAs; subsequently, dedicated Al accelerators were integrated into the chip, sharing on-chip interconnects with the CPU, GPU, ISP, and storage controller. NVIDIA's NVDLA provides an integrated open deep learning accelerator architecture, and Jetson Orin further organizes the CPU, GPU, DLA, visual accelerator, and multimedia module into a heterogeneous platform for robotics and autonomous driving [10, 11]. Qualcomm's AI Engine also adopts a form of collaboration involving the CPU, GPU, Hexagon NPU, and low-power perception modules [12]. These examples demonstrate that the performance of AI SoCs comes from the coordination of the entire platform, rather than the isolated peak performance of a single computing unit.

DOI: 10.54254/2755-2721/2026.CH37489

3.2. Applications of AI SoC in daily life

AI SoCs have been widely used in image enhancement, voice recognition, and on-device assistants in smartphones; camera perception, target detection, driver monitoring, and multi-sensor fusion in automobiles; visual positioning, path planning, and motion control in robots and drones; defect detection, condition monitoring, and predictive maintenance in industrial equipment; and low-power continuous sensing in wearable and medical terminals. The requirements for architectures vary in different scenarios: smartphones prioritize battery life and peak experience; automobiles emphasize determinism, functional safety, and redundancy; industrial systems require long-term reliability and interface compatibility; and robots demand real-time performance within the perception-decision-control loop.

3.3. Current obstacles

Similar to urban expansion where population density intensifies resource constraints, semiconductor scaling faces strict physical limitations. For the SoC, the first problem is the difficulty in heterogeneous scheduling: the performance models of CPUs, GPUs, and NPUs differ significantly, and cross-unit data migration may negate acceleration benefits. The second issue is the thermal design power limitation. Due to strict die area constraints, hot spots rapidly shift with workloads, where localized overheating triggers thermal throttling and compromises reliability. The third problem is the mismatch between the rapid evolution of AI models and the fixed nature of hardware, particularly given the continuous introduction of new operators such as Transformers, mixture-of-experts models, and generative AI. The fourth issue concerns security and trust vulnerabilities, including model theft, side-channel attacks, malicious input, and multi-tenant resource isolation. Therefore, evaluating AI SoCs requires considering performance, power consumption, thermal performance, reliability, security, and the software ecosystem from a holistic system perspective.

3.4. Future outlook of the integration of AI and SoC

With the rapid advancement of artificial intelligence technology, systems-on-chips (SoCs), serving as the "brain" of intelligent devices, are undergoing a profound transformation. Future AI SoCs are no longer merely focused on increasing computing power but are evolving towards greater intelligence, efficiency, and flexibility, as reflected in several key trends.

Firstly, the widespread adoption of edge-side generative AI will reshape user experience. Previously, complex AI tasks heavily relied on cloud servers; however, future generative AI workloads, such as text generation and image creation, will run directly on terminal devices like smartphones and personal computers. However,this poses significant challenges to chips, as these models consume a lot of memory and power. Therefore, future SoCs cannot rely solely on improving computing power; instead, they must achieve software-hardware co-optimization—such as model compression and optimized data access patterns—to ensure that powerful AI capabilities run smoothly offline, thereby preserving privacy and reducing latency [13].

Secondly, to overcome the limitations of traditional architectures, compute-in-memory (CIM) technology will become a crucial enabler. In traditional chips, data needs to be repeatedly transported between storage and calculators, which is like moving ingredients from the refrigerator to the stove and back, wasting a lot of time and energy. Future chips will attempt to embed computing units directly into storage, achieving "where data is stored, calculations are performed there", significantly reducing energy consumption and improving efficiency [14]. Concurrently,

DOI: 10.54254/2755-2721/2026.CH37489

chiplet technology is revolutionizing semiconductor manufacturing. Rather than fabricating a single, large, and costly monolithic die, engineers integrate smaller functional dies together modularly. This approach not only reduces costs but also allows different modules, such as CPUs and NPUs, to be fabricated using optimized process nodes, thereby achieving flexible system customization [15].

Furthermore, open architectures and reconfigurable computing will drive new advancements in AI SoCs. Because AI algorithms evolve rapidly, fixed hardware circuits often struggle to keep pace with software updates. Open architectures such as RISC-V allow developers to customize instructions as needed. Combined with reconfigurable computing units, this enables chips to adapt to novel AI algorithms much like software updates, significantly extending hardware lifespans [16].

Finally, with the increase in performance density of chips, intelligent thermal management and AI-assisted design have become crucial. High-performance chips generate substantial heat, prompting future SoCs to utilize AI for autonomous thermal management. Through real-time monitoring of load and temperature, they will dynamically adjust voltage and frequency, and even build a virtual "digital twin" system to predict hot spots to ensure stable operation [17]. Concurrently, leveraging AI for chip design will significantly shorten development cycles, liberating engineers from tedious layout and routing tasks so they can focus on higher-level architectural innovation [18].

3.5. From "AI on SoC" to "AI for SoC"

The integration of an SoC and AI can be further summarized into two complementary aspects. "AI on SoC" focuses on how to execute trained models efficiently on edge chips, with core issues involving operator acceleration, memory management, model compression, and deployment tools. "AI for SoC" utilizes machine learning to predict and optimize the chip itself, including design space search, performance proxy models, dynamic power control, thermal state prediction, fault detection, and resource scheduling. The former enables devices to acquire intelligent capabilities, while the latter makes the chip design and operation process more intelligent. As the system complexity increases, the two will gradually close the loop: the chip dynamically configures resources based on the current model and environment, and the model also adjusts its execution mode according to the power consumption, temperature, and bandwidth constraints of the chip.

4. Conclusion

The essence of an SoC is to integrate all computing, storage, interconnection, interface, and control functions required for a complete system onto a single platform. Artificial intelligence has changed the design goals of this platform with large-scale parallel computing, massive parameters and real-time perception. Following their integration, SoCs have progressively evolved from general-purpose systems centered on CPUs into heterogeneous architectures incorporating CPUs, GPUs, NPUs, DSPs, and dedicated modules. Their competitiveness is no longer determined by a single TOPS metric; rather, it depends on data flow, memory bandwidth, operator support, compilation tools, power consumption, thermal performance, and system reliability under real-world workloads.

In the future, AI SoCs will develop along two concurrent paths: first, enabling more complex AI models to run at lower power consumption at the edge (i.e., AI on SoC); second, utilizing AI to optimize chip design, resource scheduling, and thermal management (i.e., AI for SoC). Edge generative AI, compute-in-memory, chiplet and 3D integration, open reconfigurable architectures, and thermal digital twins are not isolated entities; instead, they will jointly drive SoCs from functional integration to collaborative intelligent system design. Therefore, the most promising

DOI: 10.54254/2755-2721/2026.CH37489

research topics often lie at the intersection of algorithms, architectures, physical constraints and application scenarios.

References

[1] Shang, R., & Ma, H. (2026, April 30). Four major "sects" besiege the "Summit of Light" of the edge and end-side AI SoC market. [News article]

[2] Xu, Q., Stok, L., Drechsler, R., Wang, X., Zhang, G. L., & Markov, I. L. (2025). Revolution or hype? Seeking the limits of large models in hardware design. arXiv preprint arXiv:2509.04905.

[3] Wolf, W. (2000). The future of system-on-chip design. *Computer, 33*(4), 162–167. https://doi.org/10.1109/2.843935

[4] Benini, L., & De Micheli, G. (2002). Networks on chips: A new SoC paradigm. *Computer, 35*(1), 70–78. https://doi.org/10.1109/2.976921

[5] Srivastava, M. B., & Chandrakasan, A. P. (2003). *Design, implementation, and optimization of systems-on-chip*. Prentice Hall.

[6] SlashGear. (2026, March 15). What is a system on a chip (SoC), and what is it used for? https://www.slashgear.com/what-is-a-system-on-a-chip-soc-and-what-is-it-used-for/

[7] Data Bridge Market Research. (2026, March 21). Market trends in relation with system-on-chip technology. https://www.databridgemarketresearch.com/reports/global-system-on-chip-market

[8] Frost, S. (2026, July 2). Lin Yongyu of Sigma Star named on Forbes China 2026 best CEOs list. *Forbes China*.

[9] PhoenixNAP. (2026, April 1). What is a system-on-a-chip? https://phoenixnap.com/kb/what-is-system-on-a-chip

[10] NVIDIA. (2023). NVIDIA deep learning accelerator (NVDLA): Open deep learning accelerator architecture. https://nvdla.org/

[11] NVIDIA. (2022). Jetson AGX Orin series technical brief. https://developer.nvidia.com/embedded/jetson-agx-orin

[12] Qualcomm. (2023). Qualcomm AI engine: Heterogeneous on-device AI computing.

[13] Navardi, M., et al. (2025). GenAI at the edge: Comprehensive survey on empowering edge devices. arXiv preprint arXiv:2502.08761.

[14] Wan, W., et al. (2022). A compute-in-memory chip based on resistive random-access memory. *Nature, 608*(7924), 504–510. https://doi.org/10.1038/s41586-022-04992-8

[15] UCIe Consortium. (2023). Universal Chiplet Interconnect Express specifications. https://www.uciexpress.org/specifications/

[16] Kumar, V., et al. (2025). Bare-metal RISC-V + NVDLA SoC for efficient deep learning inference. arXiv preprint arXiv:2503.10247.

[17] Zhang, K., Guliani, A., Ogrenci-Memik, S., & Memik, G. (2018). Machine learning-based temperature prediction for runtime thermal management across system components. *IEEE Transactions on Parallel and Distributed Systems, 29*(2), 405–418. https://doi.org/10.1109/TPDS.2017.2732951

[18] Mirhoseini, A., et al. (2021). A graph placement methodology for fast chip design. *Nature, 594*(7862), 207–212. https://doi.org/10.1038/s41586-021-03544-w