Analysis of the Principle and Configuration of AI Chip and the State-of-art Applications

Zhexi Lu

Department of Electronic Engineering, Suzhou City University, Suzhou, China

1912100139@mail.sit.edu.cn

Abstract. AI chips are specialized hardware components designed to accelerate AI operations, making them crucial for various industries. This study provides a comprehensive overview of artificial intelligence (AI) chips, their structure, principles, fabrication process, applications, limitations, and future outlooks. The article discusses the components commonly found in AI chip structures, including processing units, memory hierarchy, interconnects, and specialized hardware units. It delves into the principles guiding AI chip design, such as specialization, parallelism, low precision arithmetic, and hardware-software co-design. The fabrication process of AI chips involves intricate steps, including design, mask generation, deposition, lithography, and testing. Applications of AI chips span data centers, edge computing, autonomous vehicles, healthcare, NLP, and more. Despite their advancements, AI chips have limitations, such as energy efficiency and specificity, but the future holds promise with improvements in energy efficiency, specialized acceleration, hybrid architectures, and quantum AI chips. Overall, AI chips have transformed the AI landscape, setting the stage for continued innovation and impact in various sectors.

Keywords: AI chips, fabrication, artificial intelligence.

1. Introduction

In general, an artificial intelligence chip is a type of chip that mimics or augments human abilities using digital computers or machines run by artificial intelligence systems. Artificial intelligence (AI) chips designed to accelerate deep learning algorithms are gradually becoming more prevalent in industrial applications because conventional computing chips based on central processing units (CPUs) cannot handle the characteristics of large amounts of data, strong parallelism, and intensive computing processes in the existing deep network. The definition of an artificial intelligence chip is continually being upgraded and enhanced in academic and industrial domains as a result of advancements in AI technology. The successful implementation of the deep learning algorithm in a variety of fields, including computer vision and image recognition, speech recognition, search engine optimization, and the calculation of advertising recommendations, has significantly altered the framework of the conventional computer algorithm and made it the fundamental algorithm for the realization of artificial intelligence technology. Although the deep learning technique is straightforward and highly parallelizable, the amount of data and the size of the computing network are enormous, and the rate of computation is increasing. Realizing the importance of hardware acceleration and optimization for the advancement and use of artificial intelligence technology Therefore, the objective of current artificial intelligence chips is to maximize acceleration capability and computational power while minimizing energy usage [1].

Artificial intelligence chips, which are the foundation of the technology, are vital to the advancement of artificial intelligence. Artificial intelligence chips, often referred to as AI accelerators, AI chips, or computing cards, are expert modules that manage a lot of artificial intelligence's computational needs [2]. The implementation of neural computing at low power and cost is a growing focus of AI chips. The market for AI chips has been driven by edge computing, intelligent automation, and sensing applications [3]. At present, scientists have invented four major AI chips: GPUs, FPGAs, ASICs, and brain-like chips. They have different advantages and disadvantages, some with high computing power but low efficiency, others with balanced performance but longer R&D time, and others with more flexibility but low computing capacity. A graphics processing unit, or GPU, is a co-

processor that a computer's central processing unit (CPU) uses to offload the computationally demanding tasks necessary to generate the display on a monitor. Specialty input/output coprocessors have existed for decades, first created for mainframes and high-end workstations. GPUs, in contrast, have become widely used as a result of the growing use of personal computers and graphical operating systems and applications [4]. The FPGA chip from the Cyclone series is advanced. Additionally, the mainboard has high-capacity SDRAM or flash memory ROM. The power supply management module's inoperation can only be caused by cutting off the 5V power supply while using the normal 2.54mm connector, which can be extended. Due to its comparatively high wire and structure density, a double-layer PCB may guarantee precise, reliable, and difficult-to-err-in clock input design. The extension design is supported, the pin data is provided, and the PLL resources are reserved [5].

An ASIC is a completely unique AI chip that is designed to meet the needs of a particular application and may be adjusted at the hardware level for size, power consumption, performance, and cost. The fully customized AI chip ASIC has increasingly shown its specificity and adaptability for modern AI applications with the advancement of algorithms and integrated circuits [6]. This article focuses on the structure, principles, manufacturing processes, applications, shortcomings, and prospects for future AI chip development. In addition, a systematic summary will be made at the end.

2. Descriptions of AI Chip

AI chips, also known as AI processors or artificial intelligence chips, are specialized hardware components designed to expedite and improve computing for artificial intelligence (AI) operations. These chips are designed to meet the specific requirements of AI workloads, which frequently involve huge quantities of data processing, complicated mathematical calculations, and pattern recognition. The structure of AI chips is complex and varies depending on the specific architecture and design choices made by the chip manufacturer. However, here can provide a general overview of the key components and elements commonly found in AI chip structures (a sketch is shown in Fig. 1):

- Compute Cores: These are the primary processing units responsible for performing mathematical calculations and executing instructions. AI chips often have multiple compute cores to handle parallel processing efficiently.
- Vector Processors: Specialized units designed to perform vector operations, such as matrix multiplications, which are common in AI and neural network computations.
- Tensor Cores: These are specialized processing units optimized for tensor operations, which are fundamental to deep learning computations.
- Registers: Fast storage locations within the processing units for holding temporary data and instructions.
- Caches: Multiple levels of cache memory (L1, L2, etc.) to store frequently accessed data and reduce the need to fetch data from slower main memory.
- Main Memory (RAM): Larger memory banks used for storing data and program instructions.
- Instruction Set: A set of instructions that the processor can execute. For AI chips, this may include instructions optimized for AI and mathematical operations.
- Data Path: The part of the chip responsible for executing arithmetic and logic operations on data.
- Control Unit: Manages the flow of instructions and data within the chip, ensuring that operations are executed in the correct sequence.
- Matrix Multiply Units: Dedicated units for accelerating matrix multiplication operations commonly used in neural network computations.
- Activation Functions: Hardware support for common activation functions like ReLU (Rectified Linear Unit) used in neural networks.
- Pooling Units: Specialized units for performing pooling operations in convolutional neural networks.

- Quantization Units: Hardware designed to perform low-precision calculations for improved efficiency.
- Buses: Communication pathways that allow different components of the chip to exchange data and instructions.
- Network-on-Chip (NoC): A sophisticated interconnect that connects various components of the chip in a more scalable and efficient manner.
- Voltage Regulators: Components that regulate the voltage supplied to different parts of the chip to optimize power consumption.
- Dynamic Voltage and Frequency Scaling (DVFS): Techniques to adjust voltage and frequency based on workload to save power.
- Memory Interfaces: Connect the chip to external memory modules.
- Peripheral Interfaces: Connect the chip to external devices such as storage, sensors, and communication modules.
- Instructions for Neural Networks: AI chips may have instructions specifically designed to accelerate neural network operations.
- Tensor Processing Units (TPUs): Google's AI accelerator chip includes custom hardware for neural network inference and training.
- Clock Generators: Generate clock signals that synchronize the timing of operations within the chip.
- Power Delivery and Cooling:
- Power Delivery Network: Distributes power from the external power source to various components of the chip.
- Cooling Solutions: Heat sinks, fans, or other cooling mechanisms to dissipate heat generated during operation.

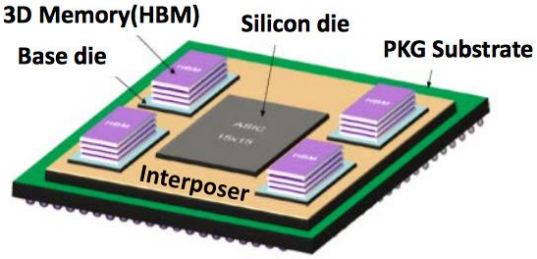


Figure 1. A sketch of the chip

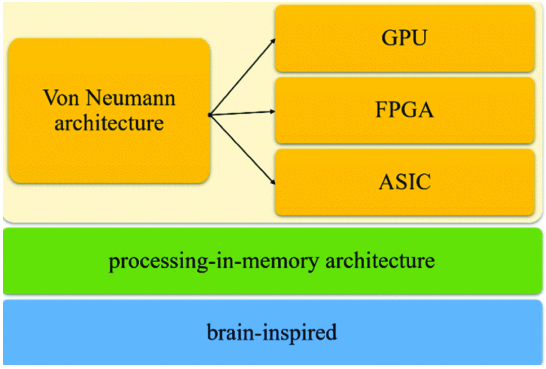


Figure 2. A sketch of architecture the AI chip

3. Principle

The principles underlying AI chips are rooted in the convergence of hardware design and artificial intelligence algorithms. These chips are engineered with specific architectural features to accelerate AI computations and tasks. Strong AI is characterized as being intelligent enough to compete with humans and having the capacity to act independently, just like humans (seen from Fig. 2). Weak AI

is the general classification for human-engineered and pre-programmed AI systems. A long-term objective of neuromorphic engineering is to replicate all of the operations of the human brain on silicon chips, which captures the transition from weak AI to strong AI [7]. For the purpose of speeding up the training and inference processes, AI chips were specifically created for AI algorithms and applications. Advanced AI chip research, which was divided into three directions from the perspective of hardware architecture, has been the driving force behind the hardware infrastructure and the environment of the machine learning platform. In the first path, the AI accelerator based on von Neumann's computer architecture is represented by the general processing unit (GPU), field programming gate array (FPGA), and application-specific integrated circuit (ASIC) [6]. Here are the fundamental principles of AI chips:

- Specialization: AI chips are purpose-built for AI workloads, concentrating on specific mathematical operations commonly found in neural networks, such as matrix multiplications and convolutions. This specialization allows them to perform these operations with exceptional efficiency.
- Parallelism: AI chips leverage parallel processing to simultaneously perform multiple computations. Neural networks inherently involve numerous parallelizable tasks, and AI chips exploit this parallelism to significantly speed up inference and training processes.
- Low Precision Arithmetic: Many AI models can tolerate lower precision calculations without compromising accuracy. AI chips use reduced-precision arithmetic, such as 8-bit or even lower, to accelerate calculations while saving power.
- Memory Hierarchy: AI chips incorporate memory hierarchies optimized for AI workloads. These hierarchies ensure quick access to frequently used data and minimize memory latency, which is critical for maintaining high throughput.
- Matrix Multiplication Acceleration: Matrix multiplication is a fundamental operation in neural networks. AI chips often feature dedicated matrix multiplication units that excel at performing these computations quickly and efficiently.
- Quantization: AI chips often quantize neural network parameters to lower bit widths, reducing memory requirements and computation complexity while preserving model accuracy.
- Hardware-Software Co-Design: AI chips are developed in tandem with AI frameworks and software libraries. Close collaboration between hardware and software teams ensures that algorithms are tailored to the chip's architecture, maximizing performance.
- Flexibility for Diverse Models: While AI chips are specialized, they are designed to support a wide range of AI models and tasks. Their programmable nature allows them to adapt to evolving AI requirements.
- Efficient Memory Usage: AI chips optimize memory access patterns and data reuse to minimize memory bottlenecks, as memory access can often be a performance bottleneck in neural network computations.
- Inference and Training Differentiation: Some AI chips are specifically designed for either inference or training tasks. Inference-focused chips prioritize low latency, while training-focused chips emphasize throughput and precision.
- Energy Efficiency: Energy efficiency is a crucial principle guiding AI chip design. These chips aim to deliver high performance while minimizing power consumption, making them suitable for various applications, including those with constrained power sources. AI models frequently contain large amounts of data, necessitating several multiply-accumulation (MAC) operations to finish a single inference or training. The majority of earlier CIM processors only supported MAC operations with fixed-point precision while having a significant improvement in energy efficiency [8].
- Scalability: AI chips are designed to scale from edge devices to data centers. This scalability ensures that the same architecture can be deployed across various platforms, maintaining consistency in AI implementations.

In summary, the principles of AI chips center on specialization, parallelism, low precision arithmetic, and optimized hardware-software co-design. By embracing these principles, AI chips unlock the potential of AI algorithms, enabling efficient and accelerated computation for a wide range of applications. The fabrication process of AI chips involves a series of intricate steps aimed at creating specialized hardware optimized for artificial intelligence tasks. Here's an overview of the fabrication process:

- Design and Architecture: The process commences with the design and architectural planning of the AI chip. Engineers create a blueprint specifying the arrangement of transistors, logic gates, memory cells, and interconnects to optimize performance for AI computations.
- Design Verification: The chip's design undergoes thorough verification using simulation and testing to identify and rectify potential design flaws or errors before moving to the manufacturing stage.
- Mask Generation: Once the design is verified, photomasks are generated. These masks serve as templates to define the layout of each layer of the chip.
- Wafer Preparation: Silicon wafers are cleaned, polished, and prepared as the base material for chip fabrication. These wafers provide a platform on which the chip's components will be constructed.
- Deposition: Thin layers of materials, such as silicon dioxide, metal, or conducting materials, are deposited onto the wafer through techniques like chemical vapor deposition (CVD) or sputtering.
- Lithography: Photomasks are used to transfer the intricate patterns onto the wafer, defining the locations of transistors, interconnects, and other components.
- Etching: Unwanted material is removed through etching processes, where chemical or physical methods selectively eliminate material to create the desired patterns.
- Implantation: Ion implantation introduces specific atoms (dopants) into the wafer's silicon structure to modify its electrical properties, essential for creating transistors.
- Annealing: Heat treatments are applied to activate dopants and repair any damage introduced during ion implantation, ensuring proper transistor behavior.
- Interconnect Formation: Metal layers are added to create the chip's wiring and interconnects, allowing different parts of the chip to communicate effectively.
- Dielectric Layers: Insulating layers are deposited between metal interconnects to prevent interference and short circuits.
- Back-End Processing: The chip undergoes additional steps involving deposition, lithography, and etching to create further metal layers and connections.
- Testing and Quality Control: At various stages, testing is conducted to ensure the components' functionality, connections, and electrical properties meet quality standards.
- Packaging: After fabrication, the wafer is cut into individual chips, which are then encapsulated in protective packaging. This packaging provides electrical connections and safeguards the chip.
- Final Testing: Each packaged chip undergoes thorough testing to validate its performance, functionality, and adherence to specifications.
- Integration: In some cases, multiple AI chips may be integrated into a larger system, alongside other components such as memory, power management, and interfaces.

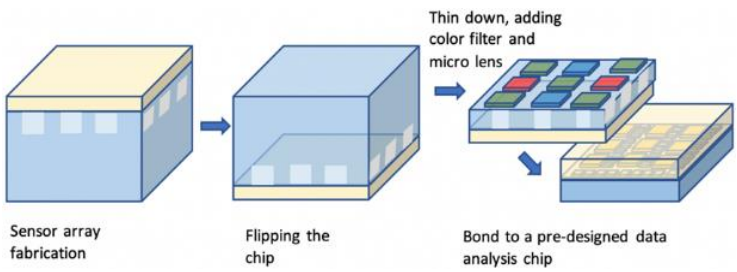


Figure 3. A sketch of fabrication

4. Fabrication and Application

The fabrication of AI chips is a meticulously executed process that demands expertise in materials science, semiconductor physics, and engineering. This process results in the creation of high-performance hardware designed to accelerate AI computations, enabling a wide range of artificial intelligence applications. A sketch is shown in Fig. 3.

AI chips are increasingly being integrated into a wide range of applications across various industries due to their ability to improve processing speed, energy efficiency, and overall performance in AI workloads. In order to achieve heterogeneous acceleration, running the DCNN network model YOLOv5s on an AI chip necessitates a number of image processing steps, including clipping, quantization, compression, and acceleration [9]. Recently complexity has been shifting from hardware to software. 32-bit CPU cores themselves are becoming a commodity. To create anything useful, applications require complex software stacks. Traditional microcontroller chip vendors had their own compilers, IDE tools, software vocabulary and libraries. This lock-in approach works for small software but does not scale for IoT where network and security code are involved. For this reason, modern software developers require commodity tools like GCC and Clang rather than closed proprietary development tools [10]. Here are some notable applications of AI chips:

Data Centers and Cloud Computing: AI chips are used in data centers and cloud computing environments to enhance the performance of AI training and inference tasks. They enable faster and more efficient processing of large datasets and complex AI algorithms, leading to improved services such as natural language processing, image recognition, and recommendation systems.

Edge Computing: AI chips are deployed at the edge of networks, closer to where data is generated, to enable real-time processing and decision-making. This is especially valuable for applications like autonomous vehicles, industrial automation, and Internet of Things (IoT) devices.

- Autonomous Vehicles: AI chips play a critical role in powering the AI algorithms required for perception, decision-making, and control in autonomous vehicles. They enable rapid processing of sensor data and help ensure safe and reliable operation.
- Healthcare: AI chips are used in medical imaging, diagnosis, and drug discovery. They accelerate tasks such as image analysis, pattern recognition, and data interpretation, leading to more accurate and efficient medical diagnoses and treatments.
- Natural Language Processing (NLP): AI chips enhance the capabilities of NLP applications, such as language translation, sentiment analysis, and chatbots. They enable faster processing of text and speech data, making interactions more natural and efficient.
- Financial Services: AI chips are employed in fraud detection, algorithmic trading, and risk assessment in the financial sector. They process vast amounts of data quickly to identify patterns and make informed decisions.
- Smart Cities: AI chips contribute to various smart city applications, including traffic management, waste management, energy optimization, and public safety. They enable real-time analysis of data from sensors and devices to improve urban efficiency.
- Agriculture: AI chips are used in precision agriculture to analyze data from sensors, drones, and satellites. They help optimize crop management, monitor soil conditions, and improve yields.
- Retail: AI chips enhance customer experiences in retail through applications such as personalized recommendations, inventory management, and cashier less checkout systems.
- Energy and Utilities: AI chips aid in optimizing energy consumption, grid management, and predictive maintenance in the energy and utilities sector. They help reduce waste and improve overall operational efficiency.
- Manufacturing: AI chips contribute to robotics, quality control, and supply chain optimization in manufacturing processes. They enable faster analysis of production data and facilitate automation.

- Security and Surveillance: AI chips enhance video analytics and object detection in security and surveillance systems. They enable real-time identification of potential threats and anomalies.
- Gaming: AI chips improve graphics rendering, physics simulations, and AI-controlled characters in gaming applications, enhancing the gaming experience.
- AI chips have made significant advancements, but they also come with limitations and exciting future possibilities.

5. Limitations and Prospects

While AI chips are more energy-efficient compared to general-purpose processors, there is still room for improvement, especially for edge devices with strict power constraints. AI chips are optimized for specific tasks, making them less versatile than general-purpose processors. This limits their application to tasks they are designed for. The design and manufacturing of AI chips require specialized knowledge, making them complex and expensive to produce. Optimizing AI chips often requires close collaboration between hardware and software developers, which can pose challenges due to different expertise domains. As AI models become larger and more complex, scaling AI chips to handle these models efficiently becomes a challenge. There's a need for standardized interfaces and frameworks to ensure interoperability between AI chips from different manufacturers.

In the future, AI chips will continue to evolve to offer even better energy efficiency, enabling longer battery life for mobile devices and reducing data center power consumption. AI chips will become more specialized for specific tasks, resulting in improved performance and reduced power consumption for those tasks. Hybrid chips combining AI accelerators with traditional processors will become more common, offering a balance between AI-specific tasks and general-purpose computing. AI chips will play a vital role in edge computing, enabling real-time processing and decision-making in devices like IoT sensors, wearables, and autonomous vehicles. AI Model Customization: Chips that allow users to customize AI models for their specific needs will become prevalent, leading to more efficient and tailored AI solutions. AI chips could incorporate features that enhance model interpretability and explainability, addressing concerns related to AI ethics and transparency. Quantum AI chips may emerge, leveraging quantum computing principles to tackle complex AI problems more efficiently. AI chips will contribute to accelerating scientific breakthroughs in fields such as medicine, materials science, and climate research. AI chips may adopt in-memory computing techniques, allowing data storage and processing to occur within the same unit for faster results. The collaboration between hardware and software developers will lead to more efficient AI algorithms that leverage the strengths of AI chips. Overall, the future of AI chips is promising, with improvements in efficiency, specialization, and adaptability that will enable a wide range of applications and drive advancements in AI technology.

6. Conclusion

In conclusion, the development of present AI chips has marked a transformative phase in the field of artificial intelligence. These chips represent a remarkable convergence of advanced hardware engineering and AI algorithmic innovation. Over the years, significant strides have been made to address challenges and optimize performance, resulting in the following key observations. In essence, the development of present AI chips has set the stage for a transformative era, enabling AI to permeate virtually every facet of our lives. Their relentless pursuit of performance, energy efficiency, and application diversity lays a strong foundation for the continued growth of AI technology, with implications that extend far beyond the realm of computation. As one move forward, the journey towards AI chip excellence promises to be both exciting and impactful.

References

[1] Wang Y, Zhang L, Zhang W. The Development and Application of Artificial Intelligence Chips. 2022 IEEE International Conference on Advances in Electrical Engineering and Computer Applications (AEECA), Dalian, China, 2022: 689-696.

[2] Cui B, Wu X, Li Y, Li J, Gu Z, Huang Y. Artificial Intelligence chip and its Technical Analysis. 2021 International Conference on Information Science, Parallel and Distributed Systems (ISPDS), Hangzhou, China, 2021: 222-225.

[3] James A P. The Why, What, and How of Artificial General Intelligence Chip Development. IEEE Transactions on Cognitive and Developmental Systems, 2022, 14 (2): 333-347.

[4] Croix J F, Khatri S P. Introduction to GPU programming for EDA. 2009 IEEE/ACM International Conference on Computer-Aided Design - Digest of Technical Papers, San Jose, CA, USA, 2009: 276-280.

[5] Zou Z. FPGA-based LED Display Technology. 2019 IEEE 4th Advanced Information Technology, Electronic and Automation Control Conference (IAEAC), Chengdu, China, 2019: 2460-2463.

[6] Gao R, Song M. Performance Comparative Analysis of Artificial Intelligence Chip Technology. 2021 2nd International Conference on Computer Engineering and Intelligent Control (ICCEIC), Chongqing, China, 2021: 149-153.

[7] James A P. Towards Strong AI with Analog Neural Chips. 2020 IEEE International Symposium on Circuits and Systems (ISCAS), Seville, Spain, 2020: 1-5.

[8] Mai Y, Wang M, Zhang C, Zhong B, Yu Z. A 1.97 TFLOPS/W Configurable SRAM-Based Floating-Point Computation-in-Memory Macro for Energy-Efficient AI Chips. 2023 IEEE International Symposium on Circuits and Systems (ISCAS), Monterey, CA, USA, 2023: 1-5.

[9] Zhang X, Meng Y, Xie X, et al. Research of YOLOv5s Model Acceleration Strategy in AI Chip. 2023 8th International Conference on Computer and Communication Systems (ICCCS). IEEE, 2023: 791-794.

[10] Hagiwara K. A two-stage-pipeline CPU of SH-2 architecture implemented on FPGA and SoC for IoT, edge AI and robotic applications. 2018 IEEE Symposium in Low-Power and High-Speed Chips (COOL CHIPS), Yokohama, Japan, 2018: 1-3.