Journal of Computer Science and Artificial Intelligence
ISSN: 3078-8242 (Print), ISSN: 3078-2309 (Online) | Vol. 5, No. 2, 2025

Research on the Evolution and Classification of Artificial Intelligence Chip Technology: A Review of Architecture Characteristics, Algorithm Adaptation, and Application Scenarios

Yinbo Hao

School of Information and Computing Science, Xi'an Jiaotong-liverpool University, Suzhou 215123, China
angstrom2006@163.com

Abstract: This paper systematically reviews the technological evolution, classification systems, and application scenarios of AI chips. Research shows AI chips have evolved from traditional general - purpose processors to a diverse ecosystem including GPUs, FPGAs, ASICs (e.g., TPUs, NPUs), and brain - inspired chips. By analyzing architectural features, algorithm adaptation, and performance limits of various chips, a multidimensional classification framework is constructed, categorizing AI chips from three dimensions: technical architecture, functional positioning, and application scenarios. The study finds general - purpose chips (e.g., GPUs) are suitable for cloud training but have low energy efficiency, specialized chips (e.g., TPUs) have excellent energy efficiency in inference but lack flexibility, and edge computing chips balance power consumption and performance. Future AI chip development will trend towards heterogeneous integration, memory - compute integration, and hardware - software co - design to tackle challenges like computational efficiency, memory wall issues, and algorithm diversity. This research offers theoretical guidance and practical references for chip selection, architectural optimization, and application deployment.

Keywords: Artificial intelligence chips, technological evolution, chip classification, architectural characteristics, algorithm adaptation, application scenarios.

1. Introduction

Artificial intelligence chips, the physical carrier and computational foundation of AI technology, are the core engine driving intelligent computing development. As deep - learning algorithms spread and computational demands grow exponentially, traditional general - purpose processors can't meet efficient computing requirements, so AI chips are evolving from general - purpose computing to specialized acceleration [1]. Currently, the market has diverse AI chip types like GPUs, TPUs, NPUs, and FPGAs, which differ in architectural design, performance characteristics, and application scenarios. Yet, neither academia nor industry has unified classification systems and evaluation standards, causing issues in chip selection, algorithm adaptation, and scenario deployment [2].

The development of AI chips has gone through multiple phases. In the early stage (pre - 2007), AI computing used general - purpose CPUs and AI chips had no independent market due to algorithm and data volume limitations. From 2007 to 2015, with the growth of high - definition video, VR, and AR gaming industries, GPUs' parallel computing was used for AI computing, improving deep - learning algorithm efficiency. After 2010, cloud computing enabled hybrid computing with many CPUs and GPUs, further promoting AI chip development. From 2015 to now, to meet rising AI computational demands, dedicated AI chips have been developed, promoting research and application with better computational efficiency and energy - consumption ratios.

Currently, the global AI chip market is growing rapidly. In 2023, the global computing power chip market reached $178.589 billion, a 1.1% year - on - year growth and accounting for 42% of the total global chip market. The Chinese market also developed rapidly, with the AI chip industry market size reaching ¥120.6 billion in 2023 and an average annual compound growth rate of 79.90% from 2019 to 2023. However, the AI chip field still faces challenges, such as balancing computing power, energy efficiency, and flexibility, and addressing memory walls and algorithm diversity.

Theoretically, establishing a chip classification framework clarifies technical boundaries and core characteristics, providing unified analytical dimensions for research. Practically, it offers design references for enterprises, hardware classification constraints for algorithm light - weighting, and accelerates AI technology scenario implementation. Different from existing research, this research uses "classification" as an anchor, extracting common patterns and differentiated issues from a classification perspective, which is unique in focus and systematics.

This paper systematically reviews the technological evolution of AI chips, proposes a multidimensional classification framework, analyzes chips' architectural characteristics, algorithm adaptability, and application scenarios, and discusses future development trends. This study fills the gap in systematic reviews of "classification - characteristics - adaptation" in the AI chip field, providing theoretical guidance and practical references for chip designers, algorithm developers, and system integrators.

2. Technological Evolution of AI Chips

The technological evolution of AI chips represents a gradual process from general-purpose computing to

specialized acceleration, driven primarily by the combined forces of algorithmic requirements, application scenarios, and semiconductor technology [3]. From a technical architecture perspective, AI chips have evolved from traditional CPU to GPU, then to FPGAs and ASICs, with recent emergence of novel architectures such as brain-inspired chips. This evolutionary process consistently revolves around three core objectives: improving computational efficiency, reducing power consumption, and enhancing flexibility.

2.1. Early Development Stage (Pre-2007)

Prior to 2007, the AI chip industry had not yet developed into a mature market. Due to limitations in algorithms, data volume, and other factors during this period, AI computing primarily relied on general-purpose CPU chips, without particularly strong market demand. Early AI algorithms were mainly rule-based systems and small-scale neural networks with relatively low computational complexity, making general-purpose CPU sufficient for application requirements. However, with the development of artificial intelligence algorithms, particularly the rise of deep learning algorithms, the demand for computing power continuously increased, gradually rendering general-purpose CPUs inadequate.

2.2. GPU Ascendancy Stage (2007-2015)

With the development of industries such as high-definition video, VR, and AR gaming, GPU products achieved rapid breakthroughs. It was discovered that the parallel computing characteristics of GPU just fit artificial intelligence algorithms and big data parallel computing requirements. GPU could improve the computational efficiency of deep learning algorithms by several tens of times compared to traditional CPU, leading researchers to begin experimenting with GPU for artificial intelligence computing [4].

Important milestones during this period include: In 1999, NVIDIA launched its first graphics processing unit, GeForce 256, initiating the era of GPU parallel computing and laying the hardware foundation for subsequent AI computing capabilities. In 2008, NVIDIA released the Tesla architecture (G80), realize unified shader model for the first time, adopting 90-nanometer process technology and supporting large-scale parallel computing. The 2010 Fermi architecture became the first complete GPU computing architecture, supporting floating-point standards and fused multiply-add instructions. The 2012 Kepler architecture optimized SM unit design, increasing the number of CUDA cores per unit to 192. The 2014 Maxwell architecture significantly improved energy efficiency ratio through SMM unit optimization and logical control efficiency.

2.3. Emergence of Specialized Chips Stage (2016-2020)

Entering the 2010s, the widespread adoption of cloud computing enabled AI researchers to perform hybrid computing utilizing large numbers of CPUs and GPUs through cloud services, further advancing the deeper application of AI chips and thereby promote the research and application of various AI chips. In 2016, NVIDIA introduced the Pascal architecture specifically for deep learning, firstly made up NVLink high-speed interconnect and HBM2 high-bandwidth memory, achieving breakthrough computational performance; concurrently, Google released the first-generation TPU, specifically optimized for machine learning, with performance 71 times that of contemporary CPUs.

Thereafter, the trend of major manufacturers developing proprietary chips accelerated: In 2017, NVIDIA's Volta architecture introduced Tensor Cores, supporting mixed-precision computing, with AI throughput increasing 12-fold compared to the previous generation; Google simultaneously iterated TPUv2/v3, focusing on cloud training and inference. During 2018-2020, Amazon launched Trainium, Microsoft released Maia100, and Meta developed MTIA chips; concurrently, Huawei's Ascend 910 emerged, marking the entry of domestic AI chips into the market.

2.4. Technological Breakthrough and Diversification Stage (2022-Present)

In 2022, the emergence of ChatGPT driven global computing power demand explosive growth. NVIDIA launched the H100 chip, with half-precision floating-point high-performance computing reaching 1979 trillion operations per second, efficiently supporting Transformer model training. Thereafter entering a period of technological intensive burst: domestic and international major manufacturers successively released their proprietary chips, NVIDIA iterated Hopper (H200) and Blackwell architectures (GB200), with the latter's computing power breaking 20,000 TFLOPS and adopting 3D packaging; Google's TPU v6 (Trillium) performance improved 4.7 times compared to v5e, with energy efficiency ratio increasing 67%; domestically, Cambricon's Siyuan 370, Hygon's DCU series accelerated adaptation to large model training, with import substitution and autonomous controllable strategies steadily progressing.

Table 1. Main Stages of AI Chip Technological Evolution

Development Stage	Time Frame	Main Characteristics	Representative Chips/Architectures	Performance Characteristics
Early Development	Pre-2007	Reliance on general-purpose CPUs	Traditional x86 CPUs	Handling complex logical operations, suitable for different data types
GPU Ascendancy	2007-2015	GPU parallel computing	NVIDIA Tesla, Fermi, Kepler, Maxwell	Powerful parallel computing capability, tens of times more efficient than CPUs
Emergence of Specialized Chips	2016-2020	Appearance of specialized acceleration chips	Google TPU, NVIDIA Pascal/Volta, Huawei Ascend 910	Specifically optimized for AI computing, significantly improved performance
Technological Breakthrough and Diversification	2022-Present	Architectural innovation and technological breakthrough	NVIDIA Hopper/Blackwell, Google TPU v6, Cambricon Siyuan 370	Computing power breakthrough, improved energy efficiency ratio, application of advanced technologies such as 3D packaging

In the evolution of AI chips, logic chips have progressively developed from simple gate circuits to complex AI accelerators. Logic chips have transformed the world, evolving from simple circuits to devices that now drive

artificial intelligence. These chips have played a key role in advancing artificial intelligence by making AI technology faster and more efficient. The global logic chip market is projected to exceed $31.5 billion by 2034, driven by growing demand for advanced processors in AI, high-performance computing, and smart devices.

An important innovation in the evolution of AI chips has been the transition from planar transistor structures to FinFET structures. Early chips used planar MOSFETs, placing the transistor gate on top of a flat channel. As engineers made transistors smaller, they faced issues such as leakage current and weakened channel control. The transition to FinFET technology addressed many of these problems. FinFET uses a thin fin-shaped channel with the gate wrapped on three sides. This design provides better current control and reduces unnecessary leakage.

In recent years, the development of AI chips has shown a trend toward diversification. In addition to traditional GPUs, FPGAs, and ASICs, new architectures such as brain-inspired chips have emerged. Brain-inspired chips are processors that mimic the neuronal structure of the human brain, featuring low power consumption, high parallelism, and adaptive learning capabilities. Although brain-inspired chips are still in the early stages of development, their potential is enormous and they are expected to become an important development direction for future AI technology.

3. Classification Framework for AI Chips

Based on the technical characteristics and application requirements of AI chips, this paper proposes a multidimensional classification framework that systematically categorizes AI chips from three core dimensions: technical architecture, functional positioning, and application scenarios. This framework aims to establish clear technical boundaries and evaluation standards, providing theoretical guidance for chip selection, algorithm optimization, and scenario deployment.

3.1. Classification by Technical Architecture

From a technical architecture perspective, AI chips can be divided into four main categories: GPU, FPGA, ASIC, and brain-inspired chips [5]. Each architecture has its unique technical characteristics and applicable scenarios.

GPU (Graphics Processing Unit) was initially designed specifically for image processing and 3D gaming. However, due to its efficient parallel computing capabilities, GPUs have gradually been applied in the AI field. Through parallel computing with thousands of small cores, GPUs can accelerate large-scale parallel computing tasks such as deep learning. GPUs are general-purpose chips that serve as mainstream acceleration chips in AI servers, possessing strong versatility and parallel computing capabilities, suitable for training and inference tasks, but with relatively high power consumption.

FPGA (Field Programmable Gate Array) is a flexible programmable hardware that can adapt to different application requirements through reprogramming. FPGAs belong to semi-custom chips, offering high flexibility and rapid iteration capabilities, suitable for small-batch diverse scenarios where algorithms are not fixed, but with high development difficulty, only suitable for fixed-point operations, and relatively expensive. Compared to GPUs, FPGAs have higher energy efficiency ratios but also higher development thresholds.

ASIC (Application Specific Integrated Circuit) is a dedicated integrated circuit designed for specific applications, featuring high performance, high efficiency, and high stability. ASICs are full-custom chips, offering optimal energy efficiency and performance, optimized for AI tasks, but with low flexibility, high R&D costs, long cycles, suitable for large-scale deployment. In the AI field, ASICs are mainly used for critical computing tasks such as intelligent speech recognition and autonomous driving.

Brain-inspired chips are processors that mimic the neuronal structure of the human brain, featuring low power consumption, high parallelism, and adaptive learning capabilities. Although brain-inspired chips are still in the early stages of development, their potential is enormous and they are expected to become an important development direction for future AI technology. Brain-inspired chips are based on neuromorphic computing, specifically Spiking Neural Networks (SNNs), expected to mature by 2025 at the earliest, with energy efficiency ratios potentially improving by 2-3 orders of magnitude compared to current chips.

3.2. Classification by Functional Positioning

According to functional positioning, AI chips can be divided into two main categories: training chips and inference chips. Training chips are mainly used for artificial intelligence algorithm training, requiring high computational precision and powerful computing support; inference chips are mainly used for artificial intelligence algorithm inference, placing greater emphasis on energy efficiency and latency.

In recent years, with the diversification of AI application scenarios, some chip types optimized for specific functions have emerged, such as:

TPU (Tensor Processing Unit): Custom-designed for deep learning tensor operations, suitable for AI training.

DPU (Data Processing Unit): Focused on data management, suitable for data movement and preprocessing in AI training/inference.

NPU (Neural Processing Unit): Simulates human neurons at the circuit level, suitable for mobile/edge real-time inference.

LPU (Language Processing Unit): Custom-designed for natural language processing scenarios, suitable for large language model real-time inference.

3.3. Classification by Application Scenario

Based on different application scenarios, AI chips can be divided into three categories: cloud chips, edge chips, and terminal chips.

Cloud chips are mainly deployed in data centers, handling training and complex inference tasks. These chips typically have powerful computing capabilities but high power consumption. Cloud environments mainly deploy training chips and inference chips, handling training and inference tasks, specifically referring to intelligent data analysis, model training tasks, and some inference tasks with high transmission bandwidth requirements.

Edge chips are deployed at the network edge, such as base stations, gateways, and other equipment, handling certain inference tasks. Edge chips need to balance computing capability and power consumption, able to independently complete data collection, environmental perception, human-computer interaction, and some inference decision control

tasks.

Terminal chips are directly integrated into terminal devices such as phones, cameras, sensors, etc., mainly completing simple inference tasks. Terminal chips typically have strict requirements for power consumption and cost, requiring high integration and optimization.

Table 2. Multidimensional Classification Framework for AI Chips

Classification Dimension	Chip Type	Main Characteristics	Advantages	Limitations	Typical Representatives
Technical Architecture	GPU	Strong parallel computing capability, good versatility	Suitable for training and large-scale parallel computing	High power consumption, high cost	NVIDIA A100, AMD MI250X
	FPGA	Reconfigurable, high flexibility	High energy efficiency, suitable for scenarios where algorithms are not finalized	High development difficulty, expensive	Xilinx Versal, Intel Agilex
	ASIC	Optimized for specific tasks	High energy efficiency, excellent performance	Low flexibility, high R&D costs	Google TPU, Huawei Ascend
	Brain-inspired chips	Mimic human brain neuronal structure	Extremely high energy efficiency, adaptive learning	Early development stage, immature	IBM TrueNorth, Intel Loihi
Functional Positioning	Training chips	High computational precision, powerful computing	Suitable for model training	High power consumption, high cost	NVIDIA H100, Google TPU v4
	Inference chips	Focus on energy efficiency and latency	High energy efficiency, fast response	Relatively low computing power	NVIDIA Jetson, Huawei Ascend 310
Application Scenario	Cloud chips	Powerful computing, supports complex models	Handles large-scale computing tasks	High power consumption, network dependency	NVIDIA DGX series, Google TPU
	Edge chips	Balance computing power and power consumption	Low latency, partial offline functionality	Limited computing power	Huawei Ascend 310, Tesla FSD
	Terminal chips	Highly integrated, extremely low power consumption	Low cost, real-time response	Limited computing power, simple functions	Apple A-series NPU, Qualcomm Hexagon

This multidimensional classification framework helps better understand the technical characteristics and application scenarios of AI chips, providing guidance for chip selection, algorithm optimization, and system design. It is important to note that these classifications are not mutually exclusive; in practical applications, integration often occur. For example, a single chip may have both training and inference capabilities, or can be used in both cloud and edge scenarios.

With technological development, the classification of AI chips continues to evolve. New chip types and architectures may emerge in the future, further enriching the classification system of AI chips. Therefore, the classification framework proposed in this paper is an open system that can be adjusted and expanded according to technological development [6].

4. Architectural Characteristics and Algorithm Adaptation Analysis of Various AI Chips

4.1. GPU Architectural Characteristics and Algorithm Adaptation

GPUs employ large-scale parallel architecture design, optimized for handling highly parallel computational tasks. Modern GPUs typically contain thousands of computing cores, utilizing multi-level memory hierarchy (including global memory, shared memory, and registers) to optimize data access. The architectural characteristics of GPUs make them particularly suitable for handling computationally intensive tasks with high parallelism, such as matrix operations and convolution operations, which are the core of deep learning algorithms.

In terms of algorithm adaptability, GPUs provide good support for mainstream neural network architectures such as CNN, RNN, and Transformer. Particularly in Transformer large model training, NVIDIA's A100 and H100 chips significantly improve training efficiency through Tensor Core support for mixed-precision computing. However, GPUs are relatively inefficient in handling sparse computations and dynamic computation graphs, and their high power consumption (e.g., A100 power consumption around 400W) limits their application in edge scenarios.

The software ecosystem is a significant advantage of GPUs. NVIDIA's CUDA ecosystem provides developers with rich tool libraries and optimization frameworks, such as cuDNN and cuBLAS, greatly reducing development difficulty. This is also an important reason why GPUs have maintained a dominant position in the AI field.

4.2. FPGA Architectural Characteristics and Algorithm Adaptation

FPGAs are based on arrays of programmable logic blocks and configurable interconnect resources, offering high flexibility and reconfigurability. FPGAs can adapt to different algorithms and application requirements through reprogramming, making them particularly suitable for application scenarios where algorithms are not yet finalized or require frequent updates.

In terms of algorithm adaptation, FPGAs can adapt to various neural network algorithms through customized programming. In 2020, Han Song's team achieved real-time AlexNet inference on FPGAs through "deep compression" technology, compressing parameters by 40 times. FPGAs perform excellently in fixed-point operations but are less efficient in floating-point operations. FPGAs are mainly suitable for small-batch, diverse inference tasks such as video

processing and signal processing.

The main challenge of FPGAs is their high development threshold, requiring hardware design expertise, and relatively long reconfiguration times. In recent years, with the development of high-level synthesis tools (HLS), the development difficulty of FPGAs has somewhat decreased, but it still remains higher than software development for GPUs.

4.3. ASIC Architectural Characteristics and Algorithm Adaptation

ASICs are dedicated chips customized for specific application scenarios, allowing for optimization at the hardware level for target algorithms, providing extremely high energy efficiency and performance. ASIC designs are typically optimized for specific computational patterns and data flows, such as matrix multiplication-accumulation operations and convolution operations [7].

In terms of algorithm adaptation, different types of ASICs are optimized for different algorithms:

TPU: Employs systolic array architecture, optimized for CNN inference tasks, with energy efficiency 30-80 times higher than CPUs, but poor adaptability to sequence models such as RNNs.

NPU: Simulates human neurons and synapses at the circuit level, processing data using deep learning instruction sets, suitable for mobile/edge real-time inference.

LPU: Custom-designed for natural language processing scenarios, suitable for large language model real-time inference.

The main advantages of ASICs are high energy efficiency, excellent performance, but low flexibility, high R&D costs, long cycles, suitable for large-scale deployment. Once algorithms change, ASICs may not adapt to new computational requirements, which is also the main risk facing ASICs.

4.4. Brain-inspired Chip Architectural Characteristics and Algorithm Adaptation

Brain-inspired chips employ neuromorphic computing architectures, mimicking the neuronal and synaptic structures of the human brain, using spiking neural networks (SNNs) for information processing. The architecture of brain-inspired chips is fundamentally different from the traditional von Neumann architecture, achieving memory-compute integration and avoiding memory wall issues.

In terms of algorithm adaptation, brain-inspired chips mainly adapt to spiking neural networks (SNNs), which are significantly different from traditional artificial neural networks (ANNs). SNNs use sparse, asynchronous spikes for communication and computation, with energy efficiency 2-3 orders of magnitude higher than traditional neural networks. However, brain-inspired chips are still in the early stages of development, with incomplete toolchains and programming models that differ significantly from traditional neural networks, limiting their widespread application.

4.5. Hybrid Architectures and System-on-Chip

In recent years, to balance flexibility, energy efficiency, and performance, hybrid architectures and system-on-chip (SoC) have gradually become trends. Hybrid architectures integrate different types of computing cores (such as CPU, GPU, NPU, etc.) on the same chip, selecting the most appropriate computing unit based on task characteristics.

System-on-Chip (SoC), as a new technology in ASIC design methodology, began in the mid-1990s. It is centered on embedded systems, based on IP reuse technology, integrating both software and hardware. It implements functions such as signal transmission, storage, processing, and I/O on a single chip, containing embedded software and the entire system's content. Due to high integration efficiency, SoC has become an inevitable trend in the development of microelectronic chips.

In terminal devices such as mobile phones and wearable devices, there are rarely independent chips; AI acceleration will be implemented by an IP on the SoC. This design can achieve higher integration, lower power consumption, and smaller size, making it very suitable for mobile devices and IoT devices.

Table 3. Comparison of Architectural Characteristics and Algorithm Adaptability of Various AI Chips

Chip Type	Computing Precision Support	Typical Computing Power Range	Energy Efficiency Ratio (TOPS/W)	Adaptable Algorithm Types	Adaptable Scenarios
GPU	FP32/FP16/FP8 /INT8	10-1000 TFLOPS	1-10	CNN, RNN, Transformer	Cloud training, inference
FPGA	INT16/INT8 /binary	1-100 TOPS	10-100	Customized algorithms	Edge inference, prototype development
ASIC	Multiple precision support	10-1000 TOPS	100-1000	Specifically optimized algorithms	Large-scale deployment, specialized scenarios
Brain-inspired chips	Spike encoding	0.1-10 TOPS	1000-10000	Spiking neural networks	Ultra-low power applications

From the table, it can be seen that different types of AI chips have significant differences in computational precision, computing power range, energy efficiency ratio, and algorithm adaptability. Selecting suitable AI chips requires comprehensive consideration of factors such as algorithm requirements, power constraints, performance requirements, and development costs.

5. Analysis of Application Scenarios for AI Chips

5.1. Cloud Data Center Scenarios

Cloud data centers are one of the most important application scenarios for AI chips, mainly used for model training and large-scale inference tasks. Cloud AI chips need to provide powerful computing capabilities, high-bandwidth memory, and high-speed interconnect technology. In cloud scenarios, GPUs currently dominate, especially NVIDIA's

series products such as A100 and H100. These chips provide powerful floating-point computing capabilities, suitable for training large-scale deep learning models.

However, with cloud computing vendors' pursuit of cost control and differentiated competition, the trend of self-developing ASIC chips is becoming increasingly apparent. Google's TPU series has been deployed on a large scale in its own cloud computing services, providing excellent performance and energy efficiency ratio. Amazon's Inferentia and Trainium chips are also used in AWS, specifically optimized for inference and training tasks. The emergence of these self-developed chips has broken GPU's monopoly in cloud computing, providing users with more choices.

The main challenges facing cloud AI chips are power consumption and heat dissipation issues. The power consumption of high-performance AI chips often reaches hundreds of watts, posing high demands on the power supply and heat dissipation systems of data centers. Additionally, memory bandwidth and capacity are also key factors limiting AI computing performance, driving the application of HBM (High Bandwidth Memory) and advanced packaging technology.

5.2. Edge Computing Scenarios

Edge computing scenarios require AI chips to balance power consumption, performance, and cost, enabling real-time processing close to the data source. Edge AI chips are increasingly being applied in non-consumer devices and occasions, such as intelligent security, ADAS/autonomous driving, smart homes, wearable smart devices, etc [8].

In the industrial field, edge AI chips have been widely applied. For example, Paifang Technology developed the Sticker series of artificial intelligence chips for industrial scenarios, covering terminal and edge computing applications. Its Tritium 103 chip focuses on one-dimensional or two-dimensional signal processing, with average power consumption below 40mW, about 80% lower than Intel's Movidius series, and energy efficiency improved by 3-4 times. It has now achieved mass production and is used in various embedded scenarios.

The characteristics of edge AI chips include: medium computing power requirements (1-20 TOPS), low power consumption (usually less than 50W), sensitivity to latency, and certain environmental adaptability. These characteristics require careful optimization of edge AI chip design in terms of architecture, process, and packaging.

In recent years, countries have been actively developing edge AI chip capabilities. In August 2025, Malaysian domestic chip design company SkyeChip officially released the country's first self-developed edge artificial intelligence chip "MARS1000", adopting advanced 6-nanometer process technology, integrating a self-developed neural network acceleration engine, capable of achieving powerful computing power of over 10 trillion operations per second (TOPS) under low power consumption conditions. This marks a historical breakthrough in Malaysia's semiconductor industry, moving toward the high end of the global value chain.

5.3. Terminal Device Scenarios

Terminal device scenarios have the most stringent requirements for AI chips, requiring extremely high energy efficiency ratios, low cost, and miniaturized size. Terminal AI chips are typically integrated into devices such as smartphones, IOT devices, cameras, etc., completing simple inference tasks such as face recognition, voice wake-up, anomaly detection, etc.

In terminal scenarios, NPUs (Neural Processing Units) have become the mainstream choice. NPUs typically employ highly optimized architectures, providing extremely high energy efficiency ratios through hardware acceleration of common neural network operations (such as convolution, pooling, activation functions, etc.). Apple's A-series chips, Qualcomm's Snapdragon series, and Huawei's Kirin series all integrate powerful NPUs, providing AI capabilities for mobile phones.

The development trend of terminal AI chips is toward higher integration, lower power consumption, and smaller size. With advances in process technology and design optimization, the computing power of terminal AI chips continues to improve while power consumption continues to decrease, enabling more and more AI applications to run on terminal devices, reducing dependence on the cloud, and improving response speed and privacy protection capabilities.

Table 4. Key Characteristic Requirements of AI Chips in Different Application Scenarios

Application Scenario	Computing Power Requirements	Power Consumption Constraints	Latency Requirements	Cost Sensitivity	Typical Representatives
Cloud Training	Extremely high (100+ TFLOPS)	Loose (300-700W)	Loose	Low	NVIDIA H100, Google TPU v4
Cloud Inference	High (10-100 TFLOPS)	Medium (100-300W)	Medium	Medium	NVIDIA T4, Google TPU v4
Edge Computing	Medium (1-20 TOPS)	Strict (5-50W)	Strict	Medium	Huawei Ascend 310, NVIDIA Jetson
Terminal Devices	Low (0.1-5 TOPS)	Extremely strict (<1W)	Extremely strict	High	Apple A16 NPU, Qualcomm Hexagon

From the table, it can be seen that different application scenarios have significantly different requirements for AI chips. Cloud scenarios pursue extreme computing power, with relatively low sensitivity to power consumption and cost; edge computing needs to balance computing power, power consumption, and latency; terminal devices are extremely concerned with power consumption and cost, with relatively low computing power requirements. These differentiated demands have also promoted the development of different types of AI chips, forming a diverse AI chip ecosystem.

6. Challenges and Future Trends

6.1. Technical Challenges

The development of AI chips faces multiple technical challenges. The most prominent one is the memory wall problem, where the performance improvement of computing units is much faster than that of memory access speed, leading to frequent waiting for data by computing units and low utilization. Another important challenge is energy efficiency

constraints, especially in edge and terminal scenarios with strict power limitations. Moreover, algorithm diversity brings adaptation challenges as different neural network architectures and computational patterns need different hardware optimization strategies.

Other technical challenges include: the maturity of software development tool-chains, chip security issues, and ensuring neural network stability. These technical challenges require joint efforts from chip designers, algorithm developers, and system optimizer to solve.

6.2. Architectural Innovation Trends

To address the aforementioned challenges, AI chip architectures are undergoing several innovations. Memory-compute integrated architecture closely integrates computing units with memory, reducing data movement and alleviating memory wall problems [9]. Chiplet technology improves yield and reduces costs by decomposing large chips into multiple small chips, supporting heterogeneous integration. Reconfigurable computing allows hardware to dynamically adjust according to workload, improving flexibility and utilization.

Heterogeneous computing is another important trend, integrating different types of processing cores (CPU, GPU, ASIC, etc.) into the same system, selecting the most appropriate computing unit based on task characteristics. Approximate computing trades reduced computational precision for improved energy efficiency and performance, particularly suitable for precision-insensitive applications. Sparse computing leverages sparsity in neural networks, skipping zero-value computations to improve computational efficiency.

6.3. Application Trends

From an application perspective, AI chips are showing the following trends: the rise of hybrid AI architectures, where terminal and cloud collaboration becomes the future. Hybrid AI chips can simultaneously meet the needs of AI processing synergy terminal and cloud, and allocate AI computing workloads appropriately according to scenarios and time, efficiently utilizing resources. Industry-specific chips are increasing, developing dedicated chips for vertical fields such as intelligent driving, medical imaging, and scientific computing.

Enhanced edge AI capabilities, with more AI computing migrating from the cloud to the edge and terminal. This trend is driven by multiple factors: reducing latency, protecting privacy, and lowering network bandwidth requirements. Hardware-software co-optimization is increasingly important, achieving overall performance optimization through co-design of algorithms and hardware.

6.4. Industrial Ecosystem Trends

In terms of industrial ecosystem, the development of AI chips shows the following trends: obvious trend of cloud vendors self-developing, with ASIC chips growing significantly. More cloud vendors such as Google, Amazon, and Meta are beginning to heavily invest in self-developing AI ASICs, promoting rapid growth in the data center custom ASIC chip market. Value of advanced packaging technology becomes apparent, with active research and development of advanced packaging processes domestically. In the context of surge computing power demand and slowing process technology advancement, advanced packaging technologies (such as Chiplet, COWOS) have become key paths to improving chip performance.

Accelerated import substitution, with Chinese AI chip enterprises developing rapidly under policy support. Domestic enterprises such as Huawei Ascend, Cambricon, and Hygon Information have achieved certain results in the AI chip field, and the trend of localization is gradually emerging. However, domestic AI chips still face challenges such as insufficient technical accumulation and synergy ecological construction.

7. Conclusion

This paper systematically studied the technological evolution, classification framework, architectural characteristics, algorithm adaptation, and application scenarios of AI chips. It found that AI chips have evolved from traditional general - purpose processors to a diverse ecosystem including GPUs, FPGAs, ASICs, and brain - inspired chips. Each type has unique characteristics and applicable scenarios, needing trade - offs among flexibility, energy efficiency, and performance.

The paper constructs a multidimensional classification framework (technical architecture, functional positioning, application scenario) to guide AI chip research and selection. General - purpose chips like GPUs are suitable for cloud training but have low energy efficiency, specialized chips like TPUs offer good energy efficiency in inference but lack flexibility, and edge computing chips balance power consumption and performance.

Future AI chip development will trend towards heterogeneous integration, memory - compute integration, and hardware - software co - design to solve challenges such as computational efficiency, memory wall issues, and algorithm diversity. Hybrid AI architectures will emerge, and terminal - cloud collaboration will be an important trend. Cloud vendors self - developing chips, advanced packaging technology, and import substitution will also shape the industry's future.

This study fills the gap in systematic reviews of "classification - characteristics - adaptation" in the AI chip field, providing optimization directions for chip designers, hardware adaptation basis for algorithm researchers, and chip selection guidelines for system integrators. Future research can focus on emerging directions like memory - compute integrated architectures, brain - inspired chips, and Chiplet technology to promote AI chip development and application.

References

[1] LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521(7553), 436-444.

[2] Jouppi, N. P., Young, C., Patil, N., Patterson, D., Agrawal, G., Bajwa, R., ... & Boyle, R. (2017). In-datacenter performance analysis of a tensor processing unit. Proceedings of the 44th Annual International Symposium on Computer Architecture, 1-12.

[3] Moore, G. E. (1965). Cramming more components onto integrated circuits. Electronics, 38(8), 114-117.

[4] Coates, A., Huval, B., Wang, T., Wu, D., Catanzaro, B., & Andrew, N. (2013). Deep learning with COTS HPC systems. International Conference on Machine Learning, 1337-1345.

[5] Sze, V., Chen, Y. H., Yang, T. J., & Emer, J. S. (2017). Efficient processing of deep neural networks: A tutorial and survey. Proceedings of the IEEE, 105(12), 2295-2329.

[6] Hennessy, J. L., & Patterson, D. A. (2019). A new golden age for computer architecture. Communications of the ACM, 62(2), 48-60.

[7] Chen, Y. H., Emer, J., & Sze, V. (2016). Eyeriss: A spatial architecture for energy-efficient dataflow for convolutional neural networks. ACM SIGARCH Computer Architecture News, 44(3), 367-379.

[8] Wang, X., Han, Y., Leung, V. C., Niyato, D., Yan, X., & Chen, X. (2020). Convergence of edge computing and deep learning: A comprehensive survey. IEEE Communications Surveys & Tutorials, 22(2), 869-904.

[9] Sze, V., Chen, Y. H., Yang, T. J., & Emer, J. S. (2017). Efficient processing of deep neural networks: A tutorial and survey. Proceedings of the IEEE, 105(12), 2295-2329.