International Journal of Science and Research Archive

eISSN: 2582-8185

Cross Ref DOI: 10.30574/ijsra

Journal homepage: https://ijsra.net/

(REVIEW ARTICLE)

Check for updates

Silicon minds: The rise of AI-powered chips

Dhruvitkumar V. Talati *

Independent Researcher, USA.

International Journal of Science and Research Archive, 2021, 01(02), 097-108

Publication history: Received on 05 December 2020; revised on 08 February 2021; accepted on 12 February 2021

Article DOI: https://doi.org/10.30574/ijsra.2021.1.2.0019

Abstract

The semiconductor industry is the fulcrum of digital revolution in the contemporary era, driving cutting-edge technologies that characterize the world today as being interconnected. Increasing demands for smart, rapid, and efficient computing continue to drive semiconductor innovation to new frontiers of possibility. In the next decade, the world semiconductor market will see explosive growth driven by disruptive technologies like artificial intelligence (AI), autonomous cars, 5G networks, and the Internet of Things (IoT). Of these, AI semiconductors, or AI chips, are proving to be a game-changer, offering peak processing and power efficiency for demanding machine learning and deep learning applications.

AI chips represent a new frontier in microprocessor design, created to speed up AI computations with unprecedented speed and efficiency. Unlike regular processors, the chips have specialized architectures, including neural processing units (NPUs) and tensor processing units (TPUs), to maximize AI workloads. Their reach spans across industries, ranging from automobile innovation supporting autonomous capability to intelligent homes with intelligent automation, robotics revolutionizing manufacturing, and AI-driven healthcare innovations. As more industries rely on AI-driven decision-making, development of semiconductor technology will play a defining role in determining the future digital economy.

This piece discusses the sudden surge of AI chips, in the wake of crucial technology innovations, competition among industries, and future trends that are shaping this new wave of semiconductor advancements. Additionally, it also points to the strategic significance of AI chips to revolutionize industries and fuel digital programs in the future.

Keywords: Semiconductors; AI Chips; Artificial Intelligence; Machine Learning; Smart Devices; Autonomous Cars; Robots; 5G; Iot; Digital Transformation

1 Introduction

Artificial Intelligence (AI) is an overnight revolution from a hypothesis to a game-changer transforming businesses and lives. At the center of this transformation is AI chips—silicon processors specifically designed to accelerate machine learning and deep learning algorithms. These are not incremental improvements of conventional processors but a complete revolution in how computers process, analyze, and decide on data. AI chips are designed uniquely to support rigorous AI operations like image recognition, voice processing, forecast analysis, and self-driving decision-making at speeds and efficiencies that standard CPUs and GPUs cannot keep up with.

1.1 The Drivers of AI Chip Innovation

The growing dependency of industries on AI-based solutions has created a demand for compute-intensive, stronger, more efficient, and scalable computing systems. Existing computing architecture is based on a great amount of

* Corresponding author: Dhruvitkumar V. Talati.

centralized processing on data centers or telecom edge processor units but is not capable of fulfilling the computing requirements of upcoming AI applications by increasing day by day. That is mostly because of the humongous amount of data created by next-generation digital interactions to be computed that needs advanced mathematics and computational algorithms.

AI chips appear as the prime facilitator of this new age, providing specialized architectures like neural processing units (NPUs), tensor processing units (TPUs), and application-specific integrated circuits (ASICs), which have the potential to improve AI processing speed by a significant margin while keeping energy consumption to a bare minimum. The likes of semiconductor leaders, cloud service providers, and AI-focused startups are making large investments in AI chip RandD, seeing their potential to enable the next generation of technological advances.

1.2 AI Chips and Robotics and Quantum Computing Future

The marriage of AI chips with other emerging technologies, including quantum computing and autonomous robotics, is going to drive their use even further. Quantum computing, which can process enormous amounts of data at one time, augments AI chips by expanding computational power exponentially. Autonomous robotics—learn, optimize, and self-directed robots—otherwise, are quickly becoming a significant domain where AI chips have a big edge. These chips allow robots to learn in real-time and to respond to a dynamic environment, make quick decisions, and function independently without being controlled by humans.

1.2.1 Abandoning the Cloud Habit

Until recently, AI computations were typically performed on faraway cloud-based data centers or high-end corporate gear, outsourcing AI model computation to third-party servers. The main reasons were:

- AI tasks consumed massive computational capabilities that standard chips were not capable of handling.
- The hardware that could sustain AI computations was costly, heavy, and energy-intensive.
- The sheer computational requirements of AI models made it unrealistic to execute them on-device.

But with technology advancements in semiconductor, AI chips are now inexpensive, compact, power-efficient, and can be used to carry out sophisticated AI computations locally. This is revolution, as with this, the AI computations will be carried out directly on the edge devices—smart home equipment, smartphones, and even driverless cars—without the constant need for access to the internet or cloud computing.

1.3 The Rise of AI at the Edge

With facilitating local AI processing, AI chips provide numerous benefits:

- Speed and Performance: AI-based applications are computed in real-time without any latency issues arising from cloud dependency.
- Increased Security and Privacy: Local storage of personal information reduces exposure to cyber-attacks and illegal access.
- Power Efficiency: AI chips are power-optimized, and hence they are best suited for battery-driven devices.
- Inter-industry Flexibility: From consumer electronics to industrial automation, AI chips are expanding possibilities in every industry.

The implications of this transformation are profound. AI chips not only transform handhelds but also find their way into autonomous cars, medical diagnostics, industrial robot control, and future robotics. By avoiding the need to send huge amounts of data to the cloud, AI chips improve real-time decision-making, opening up the possibility for quicker, more intelligent, and more efficient AI-powered systems.

This paper delves into the new landscape of AI chips, the technology behind them, and how they are going to make their mark on different industries. This paper also discusses challenges and opportunities that are in store for us as AI chips keep shaping the future of computation and digital intelligence.

2 Technology overview

Artificial Intelligence (AI) chips are arguably the most revolutionary technology in semiconductor technology, with extremely potent machine learning and deep learning solutions. There is no single definition of AI chips, but by and large, any chip that is utilized for AI computations can be termed as an AI chip. The motive behind such processors is to

tackle demanding AI workloads by leveraging optimized architectures that enable computational speed at the lowest power consumption.

Today's AI chip designs combine legacy computing architectures with new hardware and software acceleration methods. They have made it possible for AI chips to drive a broad range of applications, from autonomous cars and smart home appliances to industrial control and medical diagnostics. Unlike general-purpose processors, AI chips utilize specialized cores that are specifically designed for running parallel computations, hence making them extremely efficient in performing tasks such as pattern recognition, image processing, and natural language processing.

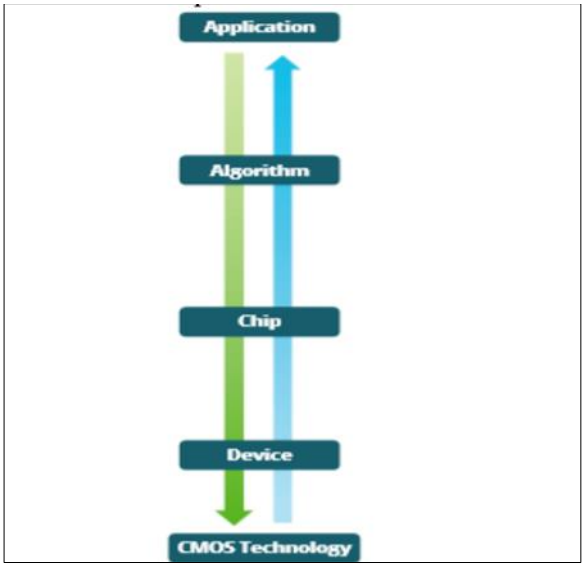

Figure 1 AI chip Technology chain

2.1 The AI Technology Stack: A Multi-Layered Solution

AI chip design is not solo; it is highly dependent in a larger AI technology stack. There are multiple dependent layers in this stack, and all of them are significant contributors to enabling AI capabilities. These layers are:

- Application Layer – Holds end-user apps that employ AI, including autonomous cars, virtual assistants, facial recognition, and smart robots.
- Algorithm and Mechanism Layer – Contains AI models and deep learning libraries that run data and perform AI operations.
- Chip and Hardware Layer – The smart component of AI processing, which performs AI jobs effectively with velocity.
- Toolchain and Software Stack – Contains AI frameworks, software development kits (SDKs), and programming interfaces that facilitate the deployment of AI models on hardware.
- Process and Material Layer – Encompasses innovations in semiconductor manufacturing, for instance, 3D stacked memory, leading lithography, and power-aware transistors that will improve AI chip performance.

AI chip technology is the middle tier of this stack that bridges theory-influenced AI breakthroughs with the real world. Application demand dictates development top-down while innovation in semiconductor process, circuit architecture, and memory technologies is the foundation support that gets AI chips rolling.

Artificial Intelligence (AI) chips are one of the most revolutionary developments in semiconductor technology. The chips are specifically designed to perform machine learning and deep learning computations in an efficient manner, offering the computational capabilities required for contemporary AI-based applications. The term AI chip is not standard, but it would typically be used to describe any semiconductor product that is specifically intended for AI-specific computations. The main goal of AI chips is to speed up deep AI workloads at low power and high efficiency.

2.2 AI Chip Architecture: Bridging the Gap Between Legacy and New Computing

New AI chip architectures combine conventional computing models with new acceleration methods in software and hardware. The marriage allows AI chips to drive a staggering number of applications, from self-driving cars and home automation devices to robots, factory automation, and imaging for medicine.

AI chips differ from general-purpose processors (like CPUs) in that they are optimized for parallel computation and thus much better than CPUs at workloads like:

- Pattern Detection – Recognizing trends in information, important for fraud detection and predictive analysis.
- Image Processing – Enhancing and processing images for medical diagnostics and facial recognition.
- Natural Language Processing (NLP) – Facilitating AI assistants, chatbots, and live translation.

2.2.1 AI chips utilize some of the architectures, such as

- Graphics Processing Units (GPUs) – Originally created for graphics computation but today highly popular in AI computations due to their capability to manage thousands of concurrent tasks.
- Tensor Processing Units (TPUs) – AI-designed chips specifically engineered to speed up deep learning computations.
- Field-Programmable Gate Arrays (FPGAs) – Programmable chips programmed for diverse AI applications.
- Application-Specific Integrated Circuits (ASICs) – AI chips custom-built to provide optimal performance and efficiency for specific applications.

3 Role of ai chip in the layers of ai

3.1 The AI Technology Stack: A Layered Approach

AI chips are not discrete hardware units, but belong to a far larger AI tech stack, that provides end-to-end AI functionality. Five are the stages in this stack which are most essential to AI processing:

3.1.1 Application Layer

The outer layer and it includes applications with AI features, which customers use day in and day out. Examples include:

- Autonomous Vehicles – Real-time sensor data is calculated using AI chips in an effort to plan self-driving buildings.
- Virtual Assistants – Voice assistants (e.g., Siri, Alexa) and chatbots powered by AI employ natural language processing to recognize and answer human queries.
- Facial Recognition Systems – AI-powered algorithms on specialized chips can reliably recognize and detect faces in security contexts.

3.1.2 Algorithm and *Mechanism Layer*

This layer includes the machine learning frameworks, deep learning algorithms, and AI models that handle massive amounts of data. CNNs, RNNs, and transformers are some technologies that run within this layer to facilitate AI vision, speech recognition, and breakthroughs in decision-making.

3.1.3 Chip and *Hardware Layer*

This is the innermost layer of AI chips, where AI computation is done by special-purpose processors. AI chips in this layer are designed for high-speed computation with low power consumption. Inventions in this layer are:

- High-speed neural processing units (NPUs)
- Edge AI chips for real-time computation on IoT devices
- Low-power AI processors for portable devices

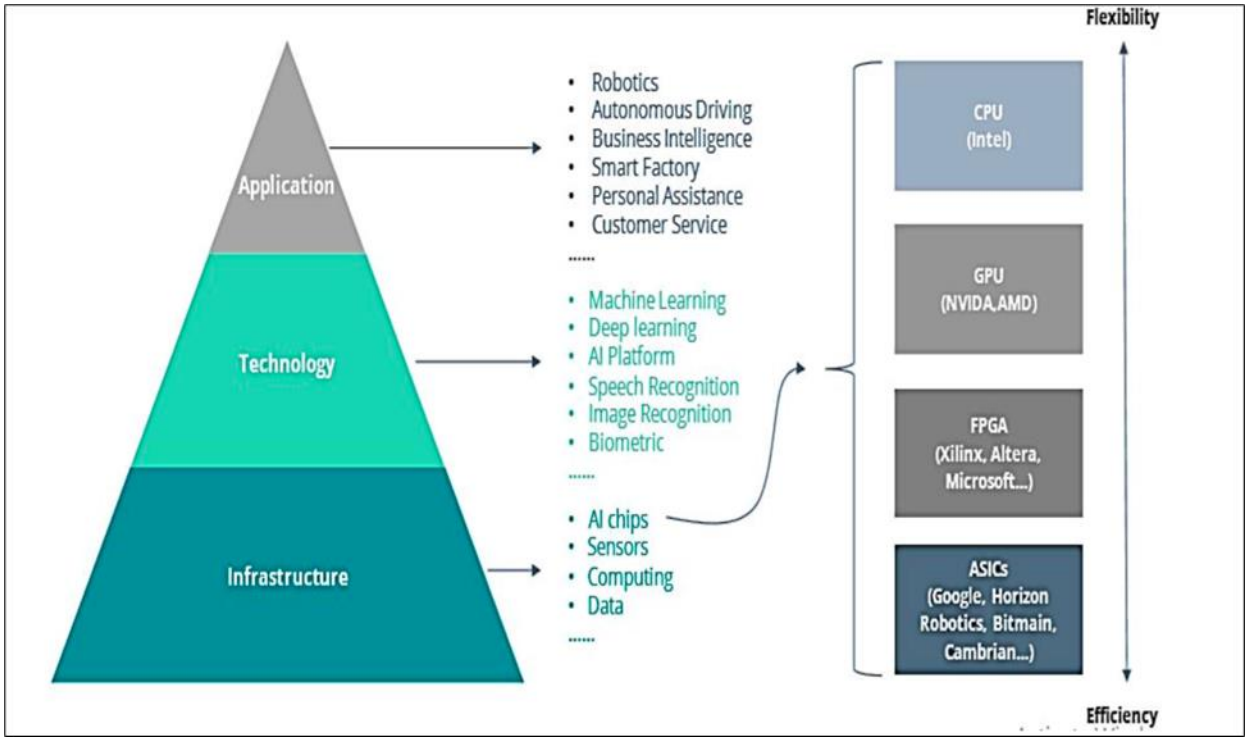

Figure 2 Role of AI chip in the layers of AI

4 Toolchain and software stack

This layer consists of the software development kits (SDKs), AI frameworks, and application programming interfaces (APIs) that facilitate successful deployment of AI models. The most popular AI frameworks include:

- TensorFlow – An open-source machine learning framework by Google.
- PyTorch – Used extensively in research and AI development.
- ONNX (Open Neural Network Exchange) – Facilitating interoperability of AI models.

5 Process and material layer

The foundation layer is focused on semiconductor manufacturing and hardware innovations that enhance AI chip performance. Advances in this layer are:

- 3D-stacked memory – Increases memory bandwidth and efficiency.
- Advanced lithography – Enables smaller but more efficient AI transistors.
- Neuromorphic computing – Mimics human brain functionality for AI applications.

5.1 AI Chips: Bridging Theory with Real-World Applications

AI chips are the essential middle connect that fulfills the bridge from theoretical AI research to practical application. Top-down demand drives AI chip development through applications of AI, but bottom-up backed by advancements in semiconductor processing technology. With this synergistic backing, AI chips continue to improve to make possible next-generation advances in autonomous technology, intelligent automation, and analytics-based AI.

5.1.1 Graphical Processing Unit (GPU)

A Graphics Processing Unit (GPU) is a specialized processor used for the sole purpose of speeding up the processing of graphics and video. They were originally developed for graphics-heavy usage like video games and digital media creation, but today they have permeated high-performance computing, primarily artificial intelligence (AI) and machine learning (ML). GPUs are designed with thousands of tiny cores that allow for parallel processing, i.e., they are able to perform multiple calculations at the same time. This feature makes GPUs extremely efficient for deep learning, where complex neural networks and big data need an enormous amount of computing power.

In AI applications, GPUs are extensively used in training and inferencing deep learning models. Training AI models entails the processing of huge amounts of data to achieve maximum performance and fine-tune parameters, which is an ideal task for the parallel computing structure of GPUs. Their flexibility and scalability have led them to become the go-to option for AI hardware, especially in cloud setups and data centers, where multiple GPUs can be linked together to accelerate AI workloads. Besides, GPUs are also finding more use in autonomous vehicles, where they compute real-time camera and sensor data to facilitate driving decisions. They are also employed in surveillance and security systems to facilitate real-time facial recognition, anomaly detection, and video analysis. Because of their flexibility, high performance, and ongoing innovation, GPUs are the most commonly used AI hardware across sectors.

5.1.2 Field Programmable Gate Array (FPGA)

A Field Programmable Gate Array (FPGA) is a semiconductor chip that may be reprogrammed for specific uses after manufacture. Unlike GPUs, whose hardware architecture is hardwired to optimize for parallel processing, FPGAs are programmable and allow programmers to have hardware tailored to their specifications. Programmability makes FPGAs particularly ideal in AI application were efficiency by power consumption, latency, or specific task must be optimized.

One of the most significant strengths of FPGAs is that they have a quicker development cycle than Application-Specific Integrated Circuits (ASICs). As they do not involve extensive manufacturing redesigns, firms are able to retune and deploy FPGA-based AI accelerators in a brief time frame, making them the preferred option for rapidly changing AI applications. Secondly, FPGAs consume less power compared to GPUs, and therefore are ideal for deployments where power consumption is a matter of primary importance, like IoT devices and edge computing. However, even with the benefits, FPGAs are still more costly than GPUs because they are programmable and niche-based.

FPGAs strike a balance between flexibility and efficiency, making them suitable for AI hardware. FPGAs are used to applications where there is a need for real-time processing, for example, telecommunications, medical imaging, and finance, where AI models must be responsive and adaptable. FPGAs are employed by AI chip makers to integrate them into systems in order to escape the expense and obsolescence of the ASIC-based strategy while achieving a high degree of optimization for certain applications.

5.1.3 Application-Specific Integrated Circuits (ASIC)

Application-Specific Integrated Circuits (ASICs) are very specialized chips with one function or application. In contrast to general-purpose GPUs and reprogrammable FPGAs, ASICs are designed from scratch for a specific task, thus being much faster, power-efficient, and performance-efficient. AI-specific ASICs are used for machine and deep learning tasks, which in turn results in the creation of specialized AI accelerators like Google's Tensor Processing Unit (TPU), Neural Processing Unit (NPU), Vision Processing Unit (VPU), and Brain Processing Unit (BPU). All these chips based on ASIC are designed particularly for different uses of AI like deep learning inference, edge AI, and computer vision.

But ASICs have some significant drawbacks. They take more time to develop and are expensive to start, as every chip needs to go through a lengthy design, testing, and fabrication process. ASICs that have already been produced can't be repurposed or rewritten, thus they are not as flexible with respect to modification of AI algorithms. If the AI model is drastically changed, an ASIC from a previous iteration might become redundant, and companies will have to pay for new hardware. In spite of this limitation, ASICs are the most efficient current AI chips to use on a large scale, especially in sectors such as cloud AI computing, autonomous systems, and high-performance AI-driven analytics.

6 Categories of AI Chips

The market for AI chips is extensively categorized in accordance with the strategies employed to implement them. There are merely two types:

6.1 Cloud-Based AI Chips

Cloud AI chips are developed for high-performance computing (HPC) tasks, in which mass-market AI models are trained and utilized via high-performance servers in data centers. Cloud AI chips are deployed on cloud computing platforms like Google Cloud AI, Amazon AWS, and Microsoft Azure AI to perform high-level AI operations like deep learning training, natural language processing (NLP), and data analysis on large scales.

One of the advantages of cloud-based AI chips is scalability. As the cloud platforms can offer almost unlimited computational resources, researchers and businesses are able to train and execute AI models that are computationally

intensive. Cloud-based AI chips are also well suited for parallel processing and are highly efficient in training neural networks and AI inference.

6.1.1 Types of cloud-based AI chips that are widespread are

- GPUs (Graphical Processing Units) – Applied to AI training and inference because of their parallelism.
- TPUs (Tensor Processing Units) – Google's proprietary AI chips that are tailored for deep learning computations.
- FPGAs (Field Programmable Gate Arrays) – Applied in cloud AI because they can be reprogrammed and are power-efficient.
- ASICs (Application-Specific Integrated Circuits) – Intended for particular AI workloads in the cloud.

Cloud AI chips drive a number of applications like autonomous vehicle simulation, AI-powered medical research, financial modeling, and large language models like ChatGPT and BERT.

6.2 Cloud-Based AI and Expanding Market for AI Chips

The biggest market for AI chips is in the cloud, where their application in data centers keeps growing. The chips are streamlining things, lowering operation costs, and streamlining infrastructure management. Of all types of AI chips, GPUs continue to be the top-selling one, with more than 30% market share. NVIDIA's GPU line has found extensive usage for deep learning applications, such as neural network training and classification, in the cloud environment. With thousands of processing cores, GPUs provide 10-100-fold application throughputs than CPUs. Consequently, they are the go-to choice for machine learning on large web and social media sites.

In addition to GPUs, specialized AI chips offer additional performance benefits. One such example is Google's Tensor Processing Unit (TPU v1), which is widely used to run AI inference workloads in the cloud, including search queries and language translation.

6.2.1 AI Training and Inference in the Cloud

Cloud AI uses big data for training models for neural networks. From large data sets of training, models are created so they can learn to recognize patterns and make predictions. After they've been trained, the model is then able to "infer" based on novel sets of data, using what it's learned to make conclusions.

The training process is highly computationally demanding and requires the usage of massive datasets over neural network models. This requires high-performance servers with increased parallel computing capacity, usually on cloud-based platforms. Owing to size and complexity of training, it is largely done in the cloud.

Inference, on the other hand, can be done either locally on edge devices or in the cloud. Training is distinct from inference because inference chips need to be powered for efficiency, low latency, and cost. Such a difference renders specialized AI chips critical to uncovering performance and resource usage trade-offs among AI applications.

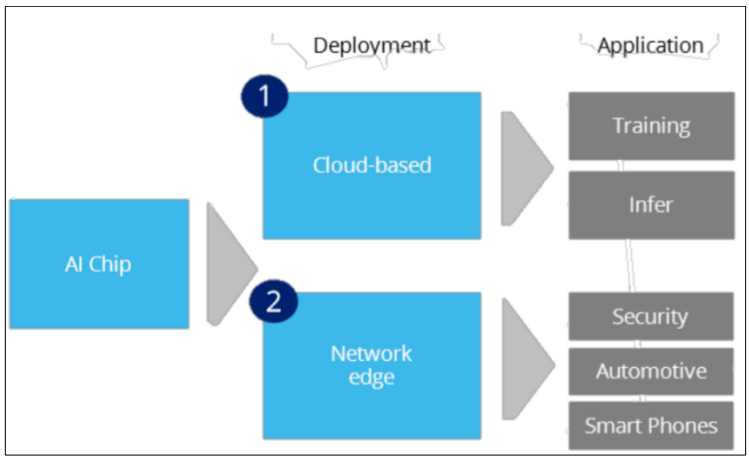

Figure 3 Categories of AI chip

6.2.2 Network Edge-Based AI Chips

Network edge-based AI chips, or edge AI chips, are engineered for on-site edge AI processing near end-user devices. In contrast to cloud-based AI chips that must be connected to remotely process AI workloads, edge AI chips process data in real time locally, hence eliminating latency and enhancing response time.

Edge AI chips are extensively used in applications where decision-making in real time is essential, for example:

- Autonomous vehicles – AI chips in autonomous vehicles scan live data from cameras, LiDAR, and sensors.
- Internet of Things devices – Smart cameras, home voice assistants, and industrial automation use Edge AI chips.
- Medical devices – AI medical wearables and diagnostic machines use edge chips for real-time processing.
- Surveillance and security – Real-time facial recognition and anomaly detection are supported by AI cameras without the need for cloud computing.

6.2.3 Some of the common edge AI chip types are

- NPUs (Neural Processing Units) – They are designed for AI inference in mobile and embedded use cases.
- VPUs (Vision Processing Units) – They are for computer vision use cases using AI.
- Low-power ASICs – They are low-power AI chips for real-time and low-power processing.

Edge AI chips are more private, less latent, and less internet-reliant, hence qualified for mission-critical deployments. They will, however, be less compute-capable AI chips compared to AI chips used in the cloud, hence qualify to be employed for inference over AI training.

The selection between edge-based and cloud-based AI chips is based on the specific AI application. Cloud-based AI chips are most suitable for high-performance training large AI models with scalability and enormous computational resources. Network edge-based AI chips, however, support real-time processing, low latency, and enhanced privacy, which is most suitable for autonomous systems, IoT devices, and mobile applications. Both play a critical role in developing AI technology across industries.

6.3 AI Chips at the Network Edge: Beyond the Cloud

AI chip deployment goes far beyond cloud data centers to a broad spectrum of network edge devices, ranging from smartphones and self-driving cars to security cameras. In contrast to AI chips in the cloud, most AI chips at the edge are inference-oriented, becoming progressively specialized to address the needs of real-time decision-making.

In some cases, cloud-trained machine learning models need to be deployed at the edge for inference because of latency, bandwidth, and privacy concerns. Power usage and cost factors are also considerations in edge AI design. For instance, in autonomous cars, real-time inference needs to happen locally, not in the cloud, to prevent adding delay by relying on the network latency.

Edge AI deployments span a great many use cases with varying computing needs. High-performance compute capability is needed by autonomous vehicles, whereas wearables need intelligent compute under extremely harsh power and budget constraints. Most edge devices will be used for inference computing once AI becomes generally adopted, so more edge inference capacity will be required. Yet, existing edge AI chips tend to fall short in satisfying local inference demands, and the market thus speeds up the roll-out and design of stronger edge AI chips to serve various use cases. The AI inference chip market is expected to register a CAGR of 40% to $2 billion by 2022.

6.4 Collaboration Between Cloud and Edge AI

Cloud-based AI processing focuses on top compute, memory bandwidth, and cost-effectiveness since it performs large volumes of data with intense parallelism and accuracy. Under such circumstances, FPGAs, GPUs, and AI-targeted silicon continue to be sought-after hardware technologies. Edge AI computing, however, focuses more on energy consumption, latency, and confidentiality.

Existing strategy to deploy AI involves training in the cloud and inferencing with the edge device. But as edge devices get stronger, increasingly the AI workloads are shifting to the edge. The future innovations can be even closer integration between cloud and edge computing, with dynamic performance-aware allocation of training and inference according to performance, latency, and power budgets. Creating this symbiotic AI ecosystem will be one of the important areas of advancement in the coming years.

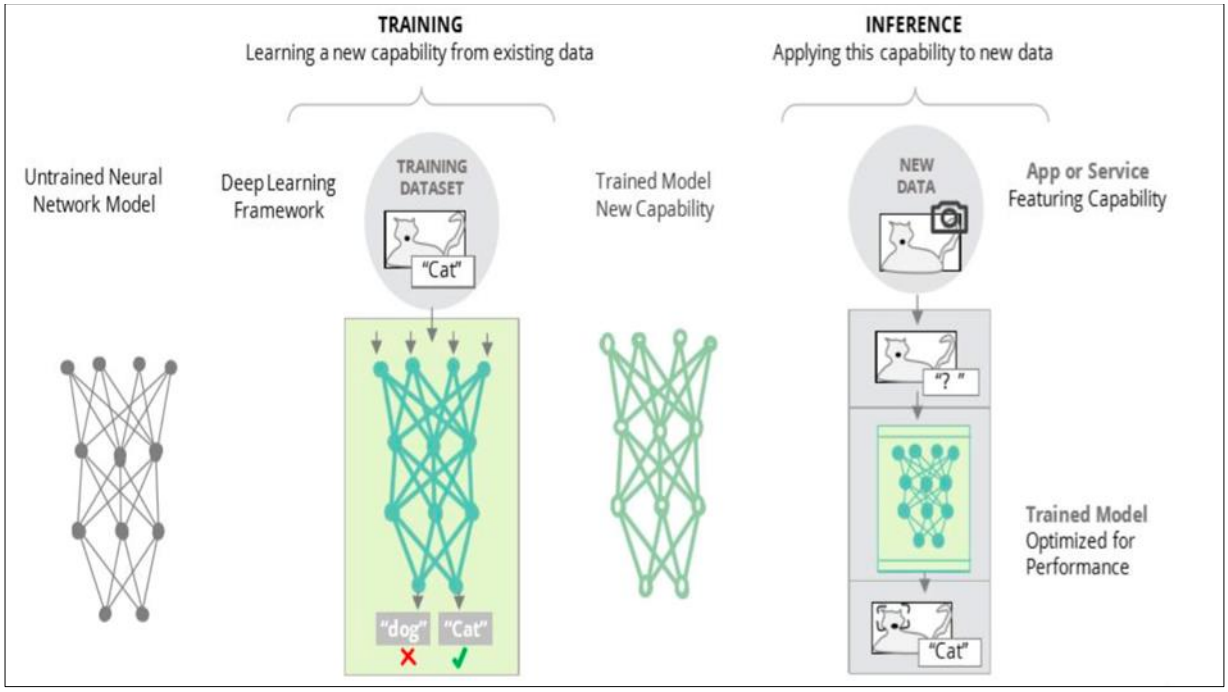

Figure 4 Phases of deep learning

7 Storage Technology for AI Chips

The performance and efficiency of AI chips rely heavily on the access and processing of data in the memory hierarchy. Since AI applications involve vast amounts of data to be processed in parallel, memory technology choice determines the power efficiency, speed, and scalability of AI chips. For the purpose of providing these high-intensity workloads, various forms of memory are used, which include AI-friendly memory, commodity memory, and on-chip (embedded) memory, each for a specific use in the AI compute environment.

AI-friendly memory is specifically tailored to provide the parallel data access performance requirements of big data and AI computing. These workloads require memory solutions to provide high-bandwidth and vast storage capacity in order to support the high rate of computation. Traditional NVM technologies are becoming increasingly difficult to scale to meet the increasing data requirements of AI models. Next-generation NVMs are therefore being researched to be a potential replacement in the near future with improved performance in bandwidth and capacity to achieve optimal AI chip efficiency. New memory technologies could potentially speed up AI workloads by large factors such that they hold much promise as an emerging trend in AI chip technology.

Commodity memory such as DRAM (Dynamic RAM) and NAND Flash are common off-chip memory for AI chips because of high capacity for storage. The memory technologies offer high storage capacity and allow data accessibility to be exploited in AI processing. 3D memory integration is amongst the key innovations in this regard, as it increases both bandwidth and capacity of commodity memory. This can be done in two fundamental ways: stacking numerous dies using Through-Silicon Via (TSV) technology or monolithic integration, wherein memory layers are constructed bottom-up. New DRAM technologies like High Bandwidth Memory (HBM) and Hybrid Memory Cube (HMC) have emerged to enhance data transfer rates such that it is easier for AI systems to conduct complicated calculations. For NAND Flash, 3D NAND technology has emerged to greatly enhance storage capacity, with significant milestones like Samsung's 96-layer 3D vertical NAND setting the boundaries for conventional flash memory.

On-chip (embedded) memory contributes significantly to AI chips by allowing rapid and effective data access inside the chip. SRAM (Static RAM) is the most prevalent on-chip memory because it is logic-circuit compatible and has the ability to enhance performance by scaling CMOS continuously. However, SRAM volatility necessitates additional non-volatile memory (NVM) support, especially for power-sensitive applications of AI chips. NOR Flash is often utilized as an on-chip NVM, but its rather slow access time and high energy expenditure may restrict overall system efficiency.

The promise of future AI chip storage lies in novel memory technologies, and these will most likely alter commodity and embedded memory uses. Foreseeable NVMs like PCM and ReRAM are being explored as SCM, which is an intermediate

between high-performance working memory and slow storage devices. These memory technologies deliver the speed required for AI workloads with higher endurance and power efficiency. Spin-Transfer Torque MRAM (STT-MRAM) is another technology with much potential as it is also being looked at as a replacement for DRAM because of its high endurance, low latency, and high read/write speeds.

Embedded NVMs also offer added benefits to the use of AI in IoT and edge devices by offering faster access times and reduced power consumption over standard NVM technology. Because most AI-powered edge devices must run within constrained power budgets, on-chip memory efficiency is essential to enable even higher performance without unnecessary overhead in power. As AI is being deployed at an increasingly larger scale, more scalable, efficient, and high-performance memory technologies will become crucial in keeping pace with future AI applications' growing computational requirements and enabling AI chips to continue doing so.

8 Benefits of AI Chips

AI chips have a number of substantial benefits, correcting some of the most crucial challenges including low connectivity, data security, processing vast volumes of data, power limitations, and low latency requirements. These processors enhance the efficiency of AI applications, which makes them more compatible with a broad array of real-world applications.

8.1 Data Security and Privacy

Data security and privacy are some of the foremost issues in AI-based applications. Companies that harvest, warehouse, and ship information to the cloud are mostly vulnerable to cyber-attacks despite having strong security policies in place to safeguard sensitive information. The nature of such risks to control becomes more essential with the growing usage of AI in sensitive fields like healthcare. For instance, intelligent speakers and AI-enabled healthcare devices are being used in hospitals where patient confidentiality is regulated.

AI chips, particularly edge AI chips, mitigate such risks by allowing for processing locally on the device, so that less of an enterprise's or a person's data has to be uploaded to the cloud. Security cameras, for instance, can be built with machine learning processors so that they can process video streams locally and understand what parts of the video stream are significant before transmitting only those to the cloud. Likewise, AI-powered voice recognition can process and understand a wider range of commands locally, thus reducing the volume of audio data to be transmitted to distant servers, while improving privacy and security.

8.2 Low Connectivity Requirements

Some cloud-connected AI-powered applications depend on the availability of a cloud connection to process and access data. There are situations where it is impossible or unpractical to have an internet connection, though. Surveillance, search and rescue, and ecological monitoring drones, for example, will typically be used far away from where there is poor or no connectivity.

In order to overcome this shortcoming, drones are now fitted with on-board AI chips that allow them to analyze data autonomously without having to connect to the cloud. AI-based drones, for instance, were used to patrol Australian beaches to track swimmers and monitor for possible threats like riptides, sharks, or crocodiles. The drones can work on their own even when there is no network connection.

8.3 Handling Big Data

The amount of data created by IoT devices, security cameras, and other connected devices is a huge challenge in storage, transmission, and analysis. For instance, security cameras across the world produce approximately 2,500 petabytes of data on a daily basis. It is expensive and complicated to transmit such massive amounts of data to the cloud for processing.

To solve this problem, AI chips embedded in edge devices locally process data prior to deciding what information is valuable enough to be stored in the cloud. For instance, vision processing units (VPUs) and low-power system-on-chip (SoC) processors facilitate real-time image processing on surveillance cameras. Rather than uploading all of the video captured to the cloud, such intelligent cameras can remove redundant data and send only essential frames for further analysis, thus lowering bandwidth and storage expenses.

8.4 Bending the Power Limits

Power consumption is among the primary disadvantages of AI-equipped devices, particularly for battery-operated devices. AI processing will take a colossal amount of processing power, and it will quickly drain the battery of a device. Low-power consuming AI chips have solved this problem by giving low-power-consuming AI processing without power wastage.

For example, AI chips based on ARM have been embedded in intelligent inhalers to monitor respiratory data like lung volume and drug delivery. AI processing takes place directly on the inhaler, and only the results of interest are sent to a coupled smartphone app. It assists medical professionals in monitoring and adjusting asthma treatment without draining battery life.

Apart from this, there have also been ongoing endeavors in research on creating deep learning models that could be executed in microcontroller units (MCUs), which are less capable and smaller than what is common for SoCs. The microcontrollers only use milliwatts or even microwatts of power, thus proving their niche for resource-scarce environments in deploying AI. Another big development in this category is TensorFlow Lite for Microcontrollers, which helps AI models easily scan and compress data on extremely low power, small chips.

9 Meeting low latency needs

AI computations performed in cloud data centers typically add latency due to the time it takes for data to travel between devices and remote servers. At best, latency is 1–2 milliseconds, but worst-case latency is tens or even hundreds of milliseconds, making real-time processing unrealistic for certain applications.

The AI chips in edge devices help to minimize latency by processing calculations at the edge, even at nanoseconds. It is especially critical for self-driving vehicles, which must process gigantic amounts of computer vision and sensor data in real time in order to make decisions at the millisecond level. Self-driving vehicles use edge AI chips and GPUs to monitor real-time input from multiple sensors, including cameras, LiDAR, and radar, and react instantly. These chips locally process environmental information to decide if the vehicle should speed up, slow down, or turn—guaranteeing the safety of passengers and seamless operation without cloud processing.

10 Conclusion

AI chips are critical in improving privacy, supporting offline AI processing, effective management of big data, power conservation, and latency minimization. With the use of AI chips at the edge, embedded devices, and cloud infrastructure, businesses are able to improve AI applications for practical use cases. AI chip technology innovation will continue to advance innovation in autonomous systems, healthcare, IoT, and smart devices, and make AI-based solutions more efficient, secure, and accessible.

Future Directions of AI Chips

The future of AI chips is not only chip-based but also on the developments in technology for memory, interconnects, and coolers. As technology in AI keeps progressing, these peripheral devices will be instrumental in improving overall system performance and efficiency.

One of the most significant problems in AI computing is high-bandwidth memory requirements. AI computations need to be computed in parallel with huge data bandwidth, thereby putting immense stress on memory systems. Therefore, the memory market will witness phenomenal opportunities by 2025 when manufacturers will struggle to make memory products faster and lower power consuming, which are particularly designed for AI workloads.

The second area of improvement is high-speed interconnects among various subsystems and devices in AI systems. With the size of AI systems increasing, the bottleneck of data transfer will become a genuine performance bottleneck. This is a great opportunity for semiconductor vendors to create next-generation interconnect solutions that will cater to the increasing data flow requirements of AI systems.

In addition, AI processors also possess a larger number of processors to achieve maximum parallelism, leading to high die sizes. This causes difficulties in power efficiency and thermal management. As power density in AI processors

increases, there is a need for heavy-duty cooling solutions to avoid overheating and maintain stability. This presents new possibilities for the package vendors to design thinner and more efficient package solutions with high thermal dissipation at lower cost. Chip packaging and thermal solution technologies will play a vital role in allowing AI hardware to operate effectively in compact form factors, including edge devices and handheld AI-based devices.

Conclusion and Future Work

Though its development has been swift, AI chip technology is new. One thing is clear, however: AI chips are a pillar of AI technology innovation and a big driver of the semiconductor industry. AI chip research today has made unprecedented advancements in machine learning acceleration, specifically in neural network-based processing, proven to excel human intelligence in nearly all computationally intensive tasks.

In the years ahead, the convergence of CMOS technology, new information technologies, and open-source AI technology will give rise to an era of unprecedented innovation. With ongoing advances in AI hardware, we can expect synergistic innovation in the space from memory to interconnects, power management, and package solutions. These innovations will not just drive the performance of AI chips but influence the next wave of AI applications, making AI solutions smarter, faster, and more efficient across markets.

Compliance with ethical standards

Disclosure of conflict of interest

No Conflict of Interest

References

[1] Chang, M.-F., Chen, A., Chen, Y., Cheng, K.-T. T., Hu, X. S., Ieong, M., Liu, Y., Van der Spiegel, J., Ling, H.-S. P., Yang, J., Yin, S., and Zhu, J. (2018). White Paper on AI Chip Technologies. Beijing Innovation Center for Future Chips (ICFC).

[2] Chou, W., Shao, J., Chung, R., Chen, L., Chen, A., and Zhou, L. (2019). Semiconductors – The Next Wave. Deloitte Touche Tohmatsu Limited (DTTL).

[3] Lee, P., Loucks, J., Stewart, D., Jarvis, D., and Arkenberg, C. (2019). TMT Predictions 2020: The Canopy Effect. Deloitte Insights.

[4] Dong, H., Shengyuan, Z., Tian, Z., Yunji, C., and Tianshi, C. (2019). A Survey of Artificial Intelligence Chips. Journal of Computer Research and Development, 56(1), 7–21. DOI: 10.7544/issn10001239.2019.20180693.

[5] Garibay, T. Y. (2018). Artificial Intelligence Chips: Past, Present, and Future.

[6] Khan, S. M., and Mann, A. (2020). AI Chips: What They Are and Why They Matter. Center for Security and Emerging Technology.