Recent Research Reviews Journal (ISSN: 2583-7079)
https://irojournals.com/rrrj

A Review on Artificial Intelligence Chip

P. Ebby Darney

Associate Professor, Department of Electrical and Electronics Engineering, Rajarajeswari College of Engineering, Bangalore, India.

E-mail: darney.pebby@gmail.com

Abstract

As chipmakers design different types of chips to enable Artificial Intelligence (AI) applications, the adoption of AI chips has increased recently. To support applications based on deep learning, AI chips have inbuilt AI acceleration and are created with a specialized architecture. One of the key drivers boosting the market's expansion is the increasing integration of AI processors in data centers. The major significance of using AI chips when compared with traditional ICs are fast computational integration and large bandwidth. This study summarizes the need of the AI chips and its functionalities and how the AI chips varies from the general ICs. Finally, a discussion on the potential AI chip initiatives are provided.

Keywords: Artificial Intelligence (AI), Field-Programmable Gate Array, Graphics Processing Units, Artificial-Intelligence Velocimetry.

1. Introduction:

The effectiveness of modern AI techniques depends on computation at a scale that was unthinkable even a few years ago. The basic of AI is computing capacity. A month of computer time may be necessary to train a top AI algorithm [1]. This huge computing capacity is provided by computer chips that not only contain the most transistors, the basic computing components that can be switched between the on (1) and off (0) states, but are also specifically designed to efficiently carry out the necessary computations for AI systems. For the same AI application to be delivered using older AI chips or general-purpose processors, it can cost tens to thousands of

DOI: https://doi.org/10.36548/rrrj.2022.1.009

times more. Therefore, cutting-edge specialized AI chips are necessary for cost-effective implementation.

The three components that fuel the development of artificial intelligence are the algorithm, processing power, and big data [2]. The training and inference of AI algorithms can be done on AI chips, which are a thousand times quicker and more effective than general-purpose CPUs. However, in work similarity to general-purpose CPUs, the AI chip is faster and efficient by integrating many tiny transistors. For the quick process and need of less energy, the small transistors are preferred over large transistors. The AI chip will increase the computing capacity depending on the number of transistors in it.

A long-term objective of AI research is to simulate the human brain in silicon and software. Moreover, the neuromorphic processors have made great progress in terms of their capacity to perform multiple numerous calculations at once and store data. They fall well short of matching the brain's capacity for conserving energy.

1.1 Need for AI Chip

The AI chips are specially made to accelerate the applications based on Artificial Neural Network (ANN). ANN uses layers of artificial neurons which are mathematical constructs followed by the function of human neurons. The next generation of mobile processors will be AI chips, since they can perform a variety of tasks more than a phone's basic features [3].

Performance: The needs of machine learning cannot be satisfied by general chips since they are not better equipped. The AI chips have an extra neural processing unit. They provide AI performance much faster and have longer battery life. AI chips provide power efficiency as well as better performance.

Technology: The AI chips are specially designed that use ML and AI technologies to construct clever devices that can replicate the human brain. AI chips use four to five times bandwidth than the standard chips, since AI applications require more bandwidth between processors in order to function properly and efficiently. This is due to their requirement for parallel processing.

Faster Computation: AI applications necessitate parallel computing capabilities in order to execute highly developed algorithms. AI chips provide superior parallel processing capabilities that are predicted to be more than ten times those of rival ANN applications [4].

1.2 AI Chip Functionalities and Architecture

One of the major challenges in AI chip designing is putting everything together. The safety and reliability requirements of the automotive industry make designing AI chips challenging. But, they are still just chips, even if with innovative processor, memory, input/output, and connection technology. To ensure that all of these new competitors in the market will be able to reach functioning silicon as soon as feasible, configurable interconnect IP can play a critical role.

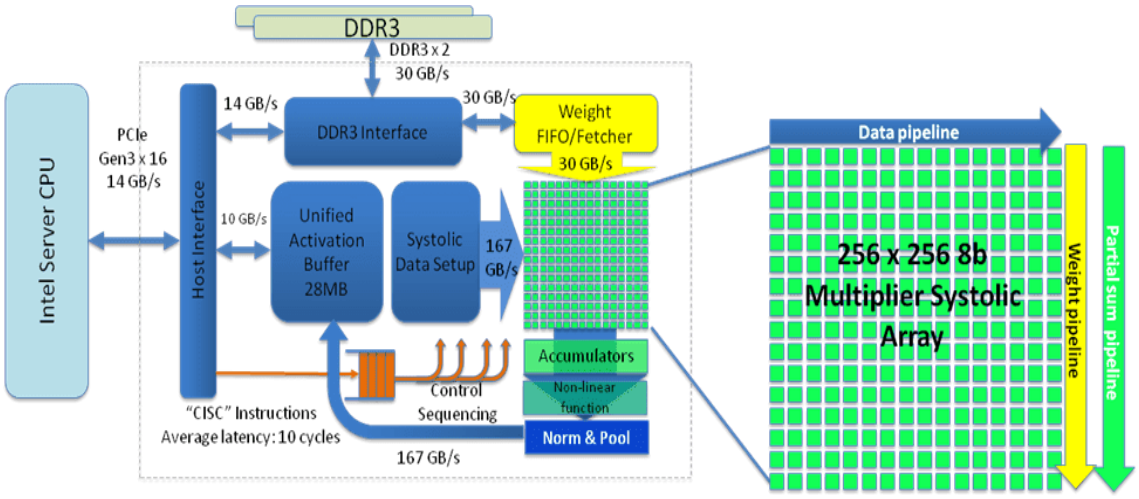

Figure 1. Architecture of AI Chip [5]

1.3. Comparison between AI Processor and General ICs:

Regular CPUs found in mobile phones and desktop computers act as engines for the devices, performing and executing whatever the phones are required to accomplish. It is the device's core hub, and it manages all the applications used on it. The future of the mobile processors lies on AI chips that can do more than just the fundamental functions of phone. Normal CPUs are placed in a smaller chip package and provide all system characteristics required to run mobile device applications. AI processors are specialized chips that use AI technology and machine learning to enable the mobile devices to mimic the human brain. AI

chip is a system that employs many processors, each with specialized capabilities. Image identification and processing become faster, allowing the smartphone to execute numerous tasks simultaneously.

Table 1. Comparison between Normal ICs and AI processor

General ICs	AI Processor
Normal CPUs are packed in a smaller chip size and are designed to handle mobile applications.	AI processors are specialized chips that use AI technology and machine learning to make the device smart.
Normal processors are not well suited to complete machine learning demands.	An AI chip is a system that refers numerous processors, each of which has a specific function.
In normal CPUs, the size, heat production and usage of energy must be reduced.	AI chips handle the programming tasks much faster than normal CPUs.

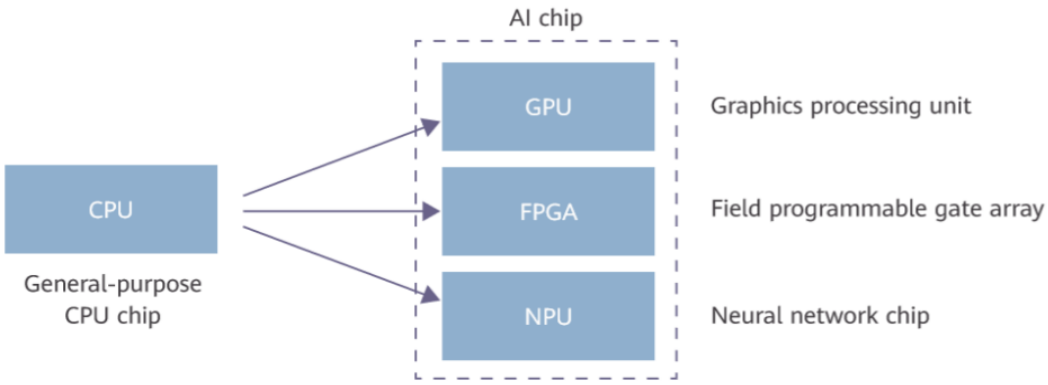

Figure 2. General ICs vs AI Chips

2. Research works related to AI chips:

Cai et al. [2], developed Artificial-Intelligence Velocimetry (AIV) technique to determine the velocity and stress domain of a human blood flow in microcirculation. The proposed AIV model

has the benefit of automated processing of experimental data using neural networks. Fetah et al., constructed a cancer-on-chip to analyze the features of cancer. On-a-chip tests will play major role in the development and approval of new cancer therapies, and it will replace the pre-clinical models [6]. Chen et al., created a Gas Sensor Array (GSA) with AI for monitoring the environment, finding gas leaks, manufacturing and storing food and beverages, and particularly illness diagnosis. The GSA is simple to use and uses low power [7]. Yoon et al., developed intellino technology to reduce the system core's workload and quickly complete the process through implementation of neural networks in hardware [8]. Gao et al., built a Network on Chip (NoC) in Communication network. It optimized intra-communication performance and energy from the following three factors: data reuse, topology, and router architecture [9].

Nwakanma et al., introduced face recognition with NeuroEdge technology using utilizing neuromorphic chip -NM500. It reduced the time required for system training and it doesn't need excessive dataset for train the systems [10]. Loke et al., created a flexible fiber with tens of meters and hundreds of interspersed with use of scalable preform-to-fiber technology. It reduced the number of necessary electrical connections to a single terminal [11]. Puri et al., developed a system based on the Internet of Things (IoT) that used several sensors to produce electricity with the utilization of Adaptive Network based Fuzzy Inference System and ANN models for predicting the production of electricity from renewable sources [12]. Kang et al., implemented an method for embedded devices to identify pedestrian images using neuromorphic chip (NM500). NM500 is more efficient than the GPU-accelerated system and it shortens the experimentation time [13]. Campero-Jurado et al., created a working prototype of a smart helmet with AI and the Industrial Internet of Things (IIoT), that keeps track of the circumstances in the workplace and near-real-time risk assessment [14].

Table 2. Comparison of Various Techniques

S.No	Reference	Technique	Applications	Outcome	Advantages
1	Cai et al. [2]	Artificial-intelligence velocimetry	1. 3D flow field 2. Microaneurysms flow in blood	Velocity and stress fields of blood flow	Neural network based automatic experimental data analysis

2	Fetah et al. [6]	Cancer-on-chip	Evaluate aspects of cancer	Development and approval of new cancer therapeutics	High throughput
3	Chen et al. [7]	Gas sensor array	Disease diagnosis and environmental monitoring	Device to be placed at home and operated by ourselves without the need of any skilled specialists	Low power consumption and ease-of-operation
4	Yoon et al. [8]	Intellino	1. Hardware implementation in neuromorphic. 2. Field-Programmable Gate Array	Distance-based AI algorithms for hardware realization	1. AI operation is performed shortly. 2. Workload of system core is reduced.
5	Gao et al. [9]	Network-on-chip	Communication network	Data reuse, Topology, Router architecture	1. Better latency. 2. Energy Saving.
6	Nwakanma et al. [10]	Neuromorphic chip-NM500	Face recognition	Extension of the Edge AI concept using neuromorphic technology	1. Doesn't required more datasets. 2. Reduce the time for training systems.
7	Loke et al. [11]	Scalable preform-to-fiber	1. Physiological monitoring. 2. Human-computer interfaces. 3. On-body machine-learning.	Introduce the tens of meters of polymeric fiber exhibiting digital sensing and memory	Reduce the number of electrical connections
8	Puri et al. [12]	ANN, Adaptive Network	1. Household appliances 2. Industrial areas	Generate electrical energy from multiple sensors	Predicts power generation from renewable resources

		based Fuzzy Inference System		using IoT based system	
9	Kang et al. [13]	Neuromorphic NM500	Software implementation	Performance and energy efficiency of a neuromorphic chip with GPU and CPU cores on embedded devices compared	Experimental time is reduced
10	Campero-Jurado et al. [14]	IIoT and CNN	Industrial and agricultural sectors	Smart helmet monitor the conditions in a working environment	Reduces the risk of accident

3. Types of AI Chips Designed for Diverse AI Applications:

AI chips are categories into four types: Application-Specific Integrated Circuits (ASIC), Field-Programmable Gate Array (FPGA), Central Processing Units (CPU), and Graphics Processing Unit (GPU) [15]**.**

Application-Specific Integrated Circuits: ASICs are silicon chips constructed for a very specific purpose. They are designed to execute a recurring function very successfully, in contrast to general-purpose chips which can carry out an infinite number of tasks less effectively. It is used to implement machine learning and AI. Because these ASICs are tailored for particular inference tasks and/or neural network structures, the logic that they implement is specifically tailored to AI computation in conventional digital logic.

Field-Programmable Gate Array: A hardware circuit containing field-programmable logic gates is known as Field Programmable Gate Array. By changing a chip's configurations, it enables users to design a unique circuit even while the device is already in use. One of the biggest performance-limiting elements in AI systems, Input/Output (I/O) bottlenecks and memory buffering, can be overcome and eliminated by FPGA.

Central Processing Units: The CPU executes fundamental mathematical, logical, controlling, and I/O activities as directed by the program's instructions. It keeps instructions, outcomes from intermediary steps and data (program), and it regulates how each component of the computer functions.

Graphics Processing Unit: GPUs are capable of performing several computations at the same time. This allows for the spread of training processes, which can greatly speed up machine learning activities. GPU can perform tasks much faster than CPU.

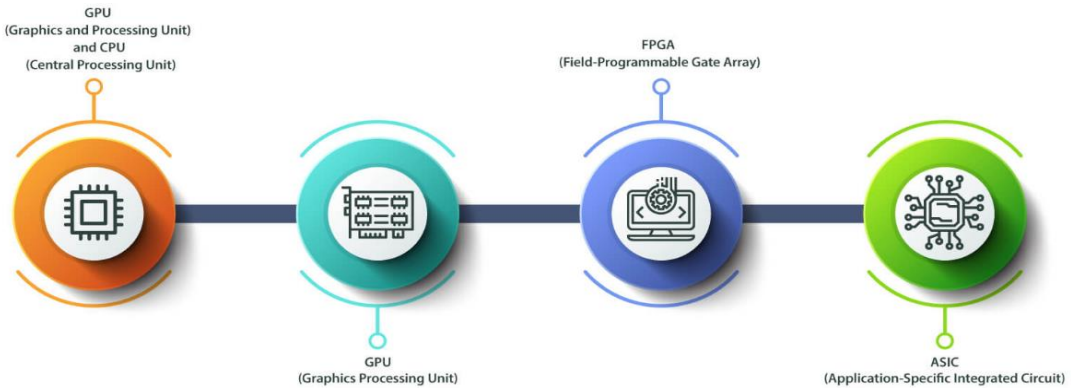

Figure 3. Diverse AI Applications

4. Discussion:

In almost every sector, artificial intelligence is influencing how people will live in the future. Emerging technologies like big data, robotics, and IoT are already primarily driven by AI chips and for the foreseeable future, it will keep innovating in the field of technology. A single chip in the size of an iPad that can transfer data thousands of times quicker than present AI chips is now being tested by at least one manufacturer and Brain Systems. This will give a possibility for the developers to test out new AI algorithms. The majority of AI chips are being developed to implement and improve versions of concepts that LeCun, Hinton, and others first proposed more than ten years ago. However, there is no reason to believe this path will result in AI that can think similar to a human being. Although AI chips currently can't meet human intelligence criteria, they are undoubtedly brilliant, and it is extremely possible that they will continue to become more brilliant in the near future. These chips will continue to use breakthroughs in semiconductor processing technology, computer designing, and SoC design to increase speed of

computation and allow next-generation AI algorithms. On the other hand, advanced memory architectures and on-chip interconnect architectures will continue to be necessary for emerging AI processors if they are to supply the proprietary hardware accelerators with the continuous flow of data needed for deep learning.

5. Conclusion:

AI chips include built-in AI acceleration and are designed with a specialized architecture to handle applications based on deep learning. The use of AI chips has a significant advantage due to their large bandwidth and quick computational integration. And the next-generation AI algorithms will be made possible by these processors by utilizing advancements in computer architecture, semiconductor processing technologies, and SoC design to increase processing power. However, the new AI processors will still require cutting-edge memory systems and on-chip interconnect designs in order to supply new proprietary hardware accelerators with the continuous stream of data necessary for deep learning.

References:

[1] Li, Bingzhen, Jiaojiao Gu, and Wenzhi Jiang. "Artificial intelligence (AI) chip technology review." In 2019 International Conference on Machine Learning, Big Data and Business Intelligence (MLBDBI), pp. 114-117. IEEE, 2019.

[2] Cai, Shengze, He Li, Fuyin Zheng, Fang Kong, Ming Dao, George Em Karniadakis, and Subra Suresh. "Artificial intelligence velocimetry and microaneurysm-on-a-chip for three-dimensional analysis of blood flow in physiology and disease." Proceedings of the National Academy of Sciences 118, no. 13 (2021): e2100697118.

[3] https://www.vyrian.com/what-is-an-ai-chip/

[4] Ye, Yuan, Zhang Yong, and Ding Han. "Research on key technology of industrial artificial intelligence and its application in predictive maintenance." Acta Automatica Sinica 46, no. 10 (2020): 2013-2030.

[5] https://semiengineering.com/artificial-intelligence-chips-past-present-and-future/

[6] Fetah, Kirsten Lee, Benjamin J. DiPardo, Eve-Mary Kongadzem, James S. Tomlinson, Adam Elzagheid, Mohammed Elmusrati, Ali Khademhosseini, and Nureddin

Ashammakhi. "Cancer Modeling-on-a-Chip with Future Artificial Intelligence Integration." Small 15, no. 50 (2019): 1901985.

[7] Chen, Zhesi, Zhuo Chen, Zhilong Song, Wenhao Ye, and Zhiyong Fan. "Smart gas sensor arrays powered by artificial intelligence." Journal of Semiconductors 40, no. 11 (2019): 111601.

[8] Yoon, Young Hyun, Dong Hyun Hwang, Jun Hyeok Yang, and Seung Eun Lee. "Intellino: Processor for embedded artificial intelligence." Electronics 9, no. 7 (2020): 1169.

[9] Gao, Wei, and Pingqiang Zhou. "Customized high performance and energy efficient communication networks for AI chips." IEEE Access 7 (2019): 69434-69446

[10] Nwakanma, Cosmas Ifeanyi, Jae-Woo Kim, Jae-Min Lee, and Dong-Seong Kim. "Edge AI prospect using the NeuroEdge computing system: Introducing a novel neuromorphic technology." ICT Express 7, no. 2 (2021): 152-157.

[11] Loke, Gabriel, Tural Khudiyev, Brian Wang, Stephanie Fu, Syamantak Payra, Yorai Shaoul, Johnny Fung et al. "Digital electronics in fibres enable fabric-based machine-learning inference." Nature communications 12, no. 1 (2021): 1-9.

[12] Puri, Vikram, Sudan Jha, Raghvendra Kumar, Ishaani Priyadarshini, Mohamed Abdel-Basset, Mohamed Elhoseny, and Hoang Viet Long. "A hybrid artificial intelligence and internet of things model for generation of renewable resource of energy." IEEE Access 7 (2019): 111181-111191.

[13] Kang, Minseon, Yongseok Lee, and Moonju Park. "Energy efficiency of machine learning in embedded systems using neuromorphic hardware." Electronics 9, no. 7 (2020): 1069.

[14] Campero-Jurado, Israel, Sergio Márquez-Sánchez, Juan Quintanar-Gómez, Sara Rodríguez, and Juan M. Corchado. "Smart helmet 5.0 for industrial internet of things using artificial intelligence." Sensors 20, no. 21 (2020): 6241.

[15] https://www.drishtiias.com/daily-updates/daily-news-analysis/artificial-intelligence-ai-chips#:~:text=AI%20chips%20are%20built%20with,the%20broader%20umbrella%20of%20AI

Author's Biography

P. Ebby Darney is working as an Associate Professor in the Department of Electrical and Electronics Engineering, RajaRajeswari College of Engineering, Bangalore, India. His area of research includes Image Processing, Artificial Intelligence, Control Systems, Radio Networks, and cloud computing.