WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Artificial Intelligence Semiconductor Technology and Development Review

Gyumin Byun [a], Daejoong Kim [b]

[a] *Department of Semiconductor Engineering, Jungwon University, Munmuro 85, Goesan eup 28024, South Korea*
[b] *Department of Semiconductor Engineering, Jungwon University, Munmuro 85, Goesan eup 28024, South Korea*

A B S T R A C T

AI semiconductors are special processors to process learning data at high speed and it can process large-scale calculations efficiently and accurately. It shows lower power consumption than traditional semiconductors as well. The AI semiconductors are important and it should be developed actively for IT technology development such as autonomous driving, AI artificial intelligence robots, etc. There are four types of AI semiconductors, including GPU, FPGA, ASIC, and Neuromorphic. In this study, the AI semiconductors technology and development by global companies are reviewed.

Keywords: Artificial Intelligence, ASIC, FPGA, Neuromorphic, Semiconductor

1. Introduction

Most artificial intelligence algorithms run on existing computing systems such as central processing units (CPUs), graphics processing units (GPUs), and field-programmable gate arrays (FPGAs). (Batra, Jacobson, Madhav, Queirolo, & Santhanam, 2019; Viswanathan, 2020) Recently, Application Specific Integrated Circuit (ASIC) of digital type or analog digital mixed-signal type is also being developed to accelerate machine learning. However, as the expansion limit of Moore's law approaches, the performance and power efficiency that can be achieved through existing expansion are decreasing. A special processor is needed to accept and process learning data in a short time, and this processor is an 'AI semiconductor.' AI semiconductors are non-memory semiconductors specialized in terms of efficiency that execute large-scale calculations required to implement AI services at ultra-high speed and ultra-power. AI semiconductors correspond to the core brain, learning data and deriving inferred results from it. (Al-Ali, Gamage, Nanayakkara, Mehdipour, & Ray, 2020; Batra et al., 2019; Esser, Appuswamy, Merolla, Arthur, & Modha, 2015) The CPU is the brain of the computer that handles all of the computer's input, output, and command processing. However, CPUs that serially process data sequentially are not optimized for AI that requires large-scale parallel processing operations. To overcome this limitation, GPU has emerged as an alternative. GPU was developed for high-end graphics processing such as 3D games, but has the characteristic of processing data in parallel, making it one of the AI semiconductors.

2. Technology of AI Semiconductor

The characteristics of AI semiconductors are as follows. FPGA is characterized by high flexibility as the hardware inside the chip can be reprogrammed according to purpose. ASIC is a custom semiconductor manufactured for a specific purpose and is characterized by high efficiency. ASIC is mainly developed by global IT companies. Neuromorphic semiconductors mimic the structure of nerve cells (neurons) and connections (synapses) that exist in the human brain. It is a next-generation AI semiconductor that is superior in performance and efficiency to previous semiconductors, but has low versatility and is still under development. (Ho et al., 2012; Huynh et al., 2022; Jeong et al., 2021; Li, Zhang, Wang, & Lai, 2020; R. Ma et al., 2022)

The AI semiconductor market size is expected to reach $12.1 billion in 2020, $18.1 billion in 2021, $24.4 billion in 2022, and $34.3 billion in 2023. This is about a 31.3% share of the entire semiconductor market. Among the semiconductors used in AI, the CPU and GPU markets, which are highly versatile, have entered the technological maturity stage. And the market is growing centered on optimized low-power, high-efficiency ASICs. (S. Ma et al., 2022; Oh, Kim, Bae, Park, & Kwon, 2020) In the future, the market share of AI semiconductors is expected to expand from semiconductors that can be used in high-performance servers such as data centers to semiconductors for devices installed in automobiles, smart phones, etc. Initially, demand for 'learning' AI semiconductors is high, but in the long term, 'inference' demand for implementing AI services based on learning data is expected to increase. Therefore, the focus of recent AI research is not only on AI algorithms, device technology, integrated systems, and architecture design, but also on efforts to overcome the computational unit-memory bottleneck of existing computers.

GPU is a processor for graphics processing, especially 3D modeling. Although GPUs are commonly used for gaming, their usage is gradually expanding due to the 4th Industrial Revolution, AI, and cryptocurrency mining. The current status of the graphics market is that NVIDIA accounts for approximately

70% and AMD accounts for 30%. In the past, CPUs were used, but the biggest reason to use GPUs now is processing speed. The CPU processes work with only about 10 cores, but the GPU processes work with hundreds of cores, so the processing speed is much faster. (Nurvitadhi, D'Souza, & Won; Pandey, Hussain, & Levy, 2020)

FPGA is a semiconductor device that contains designable logic elements and programmable internal circuitry. Logic elements that can be designed include AND OR XOR NOT. Most FPGAs include programmable logic elements plus memory elements, either simple flip-flops or complete memory blocks. FPGAs are generally slower than custom semiconductor ASICs, cannot be applied to complex designs, and consume large power. However, the development time is short, errors can be re-corrected on site, and the initial development cost is low. (Pheng & David, 2022) It provides users with a Hardware Description Language (HDL). Common hardware description languages include VHDL and Verilog. Electronic design automation tools generate technically mapped netlists. The netlist can be adapted to the actual FPGA through a process called placement and routing, usually done with software from the FPGA company. Companies such as Cadence Design Systems, Synopsis, and Celoxica use System C as a high-level language approach to accelerate the FPGA design cycle than using traditional hardware description languages. Figure 1 shows the basic structure of FPGA.

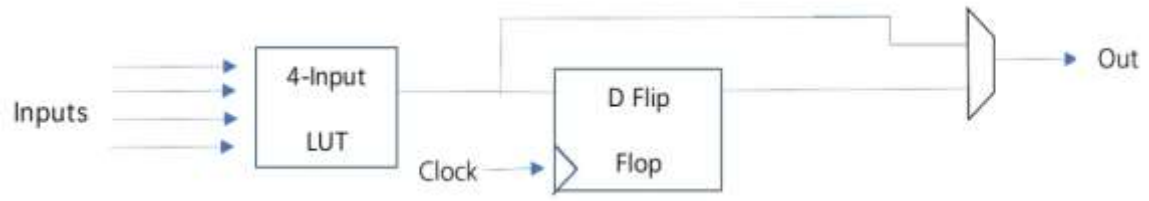

Fig. 1 - Basic structure of FPGA

ASIC is an application-specific semiconductor. It is an integrated circuit made for a specific application field and a special function of a device, and is a semiconductor used for a specific purpose. It is classified into two types: fully custom IC and semi-custom IC. It has characteristics such as small size, high performance, ultra-low power, ultra-low noise, improved reliability, cost reduction, IP safety, long-term supply, and customized manufacturing.

Neuromorphic is a technology that combines the principles of brain science. It is designed according to the operating principles of neuron neural networks, and processes data by controlling the connections and synapse strength between neurons, similar to the brain's learning and memory functions. (Qi & Yao, 2021; Saeed et al., 2023; Senoner, Netland, & Feuerriegel, 2022; Sharad, Augustine, Panagopoulos, & Roy, 2012; Shi et al., 2015; Viswanathan, 2020) While traditional semiconductors process tasks sequentially, neuromorphic performs simultaneous calculations and information processing. Currently, the technology for implementing neuromorphic systems in hardware can be divided into a method of implementing hardware using only traditional silicon transistors and a method of implementing hardware through the fusion of silicon transistors and next-generation neuromorphic devices such as memristors. Once neuromorphic hardware composed of neurons and synaptic circuits is implemented, various neural net algorithms can be transplanted into neuromorphic hardware to imitate various cognitive functions of the human brain. There are various algorithms such as a simple perceptron algorithm, an artificial neural network (ANN) algorithm based on back-propagation learning, and a spiking neural network (SNN) based on spike signals. Even today, many new brain-mimicking neural network algorithms continue to be developed.

3. Development of AI semiconductors

In the GPU field, NVIDIA is leading the market by launching new high-performance GPUs such as A100 and H100, and NVIDIA's GPUs are also used in ChatGPT and Microsoft Azur. AMD and Intel are also producing GPUs and are chasing NVIDIA. The table 1 below shows the development status and functions of each artificial intelligence GPU.

In the FPGA field, AMD is leading the market by acquiring Xilinx, which has the largest market share, and Intel is catching up by acquiring Altera. Microsoft launched Catapult and applied it to its search engine, Bing, while also applying the function of generative artificial intelligence to its Office program to provide new services.

Table 1 - Development status and functions of each artificial intelligence GPU

		NVIDIA					Google
		A2	L4	A30	A100	H100	TPUv4
Usage		**Inference**	**Inference**	**Learning**	**Learning**	**Learning**	**Learning**
Release		21.11	23.3	21.4	20.5	23	21.5
Computation performance	FP16/ BF16	-	-	330	624	1513	275
	INT8	72	485	661	1,248	3,026	-
Power Consumption (W)		40~60	72	165	30	300~350	170

Looking at the ASIC field, in addition to development and production at existing semiconductor companies, big tech companies such as Google, Tesla, and Apple are also developing and utilizing AI semiconductors for use in their products. Representative semiconductors such as Tesla's FSD (Full Self Drive) chip and Apple's processor A15 are being developed and applied. In the memory semiconductor field, semiconductors are being developed using the PIM (Processor In Memory) method, which implements calculation and parallel processing functions within the memory. This can improve computation and energy efficiency by improving the data movement structure between CPU, GPU processors and memory. Samsung Electronics and SK Hynix are also developing PIM.

In the early stages of research, IBM, Intel, and Qualcomm are participating in the early market by launching commercial products. Furiosa AI has launched FPGA-based Warboy and is applying it to Kakao Enterprise and e-Popsoft, while Rebellion and Sapien have launched ASIC-based AI semiconductor products for servers. These companies' AI semiconductors show high performance in terms of computational speed and performances compared to existing GPUs, and are also highly competitive in price. Neuromorphic research is currently being actively conducted worldwide. The table2 shows the status of artificial intelligence processor development by companies. NVIDIA products that support a level of computational efficiency suitable for building an AI system include A2, L4, A30, and A100, and H100 is a product that more than doubles the performance of previous products. Nvidia A100, which is known to be used to build ChatGPT, mainly performs matrix operations. Google unveiled TPUv4, which is equipped with about 4,000 AI semiconductors and can achieve speeds more than twice as fast as existing TPUs. As a way to show computational performance, both FP16 and BF16 are numeric formats and are a way to express floating point numbers processed in computer operations. FP16 and BF16 are floating point formats that help models run faster and save memory, but have lower accuracy. BF16 is a 16-bit floating point format and is widely used in the field of artificial intelligence. FP16 is a 16-bit floating point format and uses less memory, so it is widely used in the field of deep learning. It is less accurate than the 32-bit floating point format, but has lower memory requirements, making it useful for model training. Therefore, it is possible to use FP32 during model training and FP16 during the inference step to simultaneously increase accuracy and computation speed.

Table 2 - Status of artificial intelligence processor development by companies

Company	**Processor**			
	GPU	**FPGA**	**ASIC**	**Neuromorphic**
Google	X	X	TPU	X
Amazon	X	X	Inferentia Trantium	X
Apple	X	X	A15Bionic	X
Tesla	X	X	FSD	X
intel	H3CXG310	Agilex	NervanaNNP	Loihi
AWD	InstinctM100	Alveo	X	X
NVIDIA	VersalAICore A100,H100	X	Xavier	X

4. Conclusion

AI semiconductors are special processors needed to process learning data at high speed. AI semiconductors can process large-scale calculations quickly and accurately. The reason why AI semiconductors are currently widely used is because they have faster processing speed and lower power consumption than existing semiconductors, and they are widely used as GPUs for processing graphics. AI semiconductors are a task for the IT industry that must be developed to actively utilize future technologies – autonomous driving, AI artificial intelligence robots, etc. There are a total of four types of AI semiconductors, including GPU, FPGA, ASIC, and Neuromorphic. As large global companies are focusing on the development of AI semiconductors, the potential for future development appears to be high..

Acknowledgements

This research was supported by "Regional Innovation Strategy (RIS)" through the National Research Foundation of Korea(NRF) funded by the Ministry of Education(MOE)(2021RIS-001).

References

Al-Ali, F., Gamage, T. D., Nanayakkara, H. W., Mehdipour, F., & Ray, S. K. (2020). Novel casestudy and benchmarking of AlexNet for edge AI: From CPU and GPU to FPGA. Paper presented at the 2020 IEEE Canadian Conference on Electrical and Computer Engineering (CCECE).

Batra, G., Jacobson, Z., Madhav, S., Queirolo, A., & Santhanam, N. (2019). Artificial-intelligence hardware: New opportunities for semiconductor companies. McKinsey and Company, 2.

Esser, S. K., Appuswamy, R., Merolla, P., Arthur, J. V., & Modha, D. S. (2015). Backpropagation for energy-efficient neuromorphic computing. Advances in neural information processing systems, 28.

Ho, M.-H., Ai, Y.-Q., Chau, T. C.-P., Yuen, S. C., Choy, C.-S., Leong, P. H., & Pun, K.-P. (2012). Architecture and design flow for a highly efficient structured ASIC. IEEE transactions on very large scale integration (VLSI) systems, 21(3), 424-433.

Huynh, P. K., Varshika, M. L., Paul, A., Isik, M., Balaji, A., & Das, A. (2022). Implementing spiking neural networks on neuromorphic architectures: A review. arXiv preprint arXiv:2202.08897.

Jeong, C., Myung, S., Huh, I., Choi, B., Kim, J., Jang, H., . . . Jang, W. (2021). Bridging TCAD and AI: Its application to semiconductor design. IEEE Transactions on Electron Devices, 68(11), 5364-5371.

Li, Z., Zhang, Y., Wang, J., & Lai, J. (2020). A survey of FPGA design for AI era. Journal of Semiconductors, 41(2), 021402.

Ma, R., Georganas, E., Heinecke, A., Gribok, S., Boutros, A., & Nurvitadhi, E. (2022). FPGA-based AI smart NICs for scalable distributed AI training systems. IEEE Computer Architecture Letters, 21(2), 49-52.

Ma, S., Wu, T., Chen, X., Wang, Y., Tang, H., Yao, Y., . . . Wan, J. (2022). An artificial neural network chip based on two-dimensional semiconductor. Science Bulletin, 67(3), 270-277.

Nurvitadhi, E., D'Souza, R., & Won, M. Real performance of FPGAs tops GPUs in the race to accelerate AI. Intel-White Paper, intel. com.

Oh, K., Kim, S., Bae, Y., Park, K., & Kwon, Y. (2020). Trend of AI Neuromorphic Semiconductor Technology. Electronics and Telecommunications Trends, 35(3), 76-84.

Pandey, B. K., Hussain, D. A., & Levy, J. (2020). AI and FPGA-Based IoT Architectures, Models, and Platforms for Smart City Application IoT Architectures, Models, and Platforms for Smart City Applications (pp. 94-106): IGI Global.

Pheng, M. S. K., & David, L. G. (2022). Artificial intelligence in back-end semiconductor manufacturing: A case study. Paper presented at the 2022 IEEE International Conference on Distributed Computing and Electrical Circuits and Electronics (ICDCECE).

Qi, L., & Yao, K. (2021). Artificial intelligence enterprise human resource management system based on FPGA high performance computer hardware. Microprocessors and Microsystems, 82, 103876.

Saeed, U., Tunio, I. A., Hussain, M., Memon, F. A., Hoshu, A. A., & Hussain, G. (2023). A Review of Structural Testing Methods for ASIC based AI Accelerators. IJCSNS, 23(1), 103.

Senoner, J., Netland, T., & Feuerriegel, S. (2022). Using explainable artificial intelligence to improve process quality: Evidence from semiconductor manufacturing. Management Science, 68(8), 5704-5723.

Sharad, M., Augustine, C., Panagopoulos, G., & Roy, K. (2012). Proposal for neuromorphic hardware using spin devices. arXiv preprint arXiv:1206.3227.

Shi, L., Pei, J., Deng, N., Wang, D., Deng, L., Wang, Y., . . . Song, S. (2015). Development of a neuromorphic computing system. Paper presented at the 2015 IEEE international electron devices meeting (IEDM).

Viswanathan, S. M. (2020). AI Chips: New Semiconductor Era. International Journal of Advanced Research in Science, Engineering and Technology, 7(8), 14687-14694.