International Journal of Internet, Broadcasting and Communication Vol.12 No.1 90-94 (2020)
http://dx.doi.org/10.7236/IJIBC.2020.12.1.90

IJIBC 20-1-12

Memory Design for Artificial Intelligence

Doosan Cho

Professor, Department of electronic engineering, Sunchon National University, Korea
dscho@scnu.ac.kr

Abstract

Artificial intelligence (AI) is software that learns large amounts of data and provides the desired results for certain patterns. In other words, learning a large amount of data is very important, and the role of memory in terms of computing systems is important. Massive data means wider bandwidth, and the design of the memory system that can provide it becomes even more important.

Providing wide bandwidth in AI systems is also related to power consumption. AlphaGo, for example, consumes 170 kW of power using 1202 CPUs and 176 GPUs. Since more than 50% of the consumption of memory is usually used by system chips, a lot of investment is being made in memory technology for AI chips. MRAM, PRAM, ReRAM and Hybrid RAM are mainly studied. This study presents various memory technologies that are being studied in artificial intelligence chip design. Especially,MRAM and PRAM are commerciallized for the next generation memory. They have two significant advantages that are ultra low power consumption and nearly zero leakage power. This paper describes a comparative analysis of the four representative new memory technologies.

Keywords: *Memory system, Artificial intelligence, Power consumption, System design, Memory bandwidth*

1. Introduction

Artificial intelligence (AI) [1] applications have different characteristics from existing applications. Innovative memory technology is required in both portable devices and data centers. New memory technologies must be able to optimize scheduling for higher memory densities and data access patterns. It must also be able to provide high energy efficiency while minimizing data exchange with cloud data centers.

According to the demand of the AI market, NAND flash [2], 3D XPoint (Intel's Optane) [3], Phase-Change Memory (PCM) [4], Resistive Random Access Memory (ReRAM) [5], and Magnetic Random Access Memory (MRAM) [6] are that significant developments have been made for the next generation of memory technology in AI application domain.

The energy-efficient, durable and non-volatile memory all help make model training and reasoning more efficient on mobile devices. Additional benefits include potential improvements in stability and throughput. This improvement in memory technology helps to minimize the communication overhead with the data cloud

Manuscript Received: December. 13, 2019 / Revised: December. 20, 2019 / Accepted: December. 27, 2019
Corresponding Author: dscho@scnu.ac.kr
Tel: +82-61-750-3577, Fax: +82-61-750-3570
Professor, Department of Electronic Engineering, Sunchon National University, Korea

server system.

In particular, some new memory technologies offer clear advantages in terms of optimization characteristics for various AI applications. ReRAM and PCM offer advantages for inferential applications due to their superior speed (compared to flash), density and non-volatility. MRAM offers similar advantages as ReRAM and PCM. It is also extremely durable to compete with and complement Static Random Access Memory (SRAM), as well as perform flash replacement functions. Table 1 shows that three representative memory technologies and its characteristics. Table 2 describes properties (such as latency, endurance, retention) of the five most important memory technologies.

Table 1. Memory classification

	How it works	Characteristics
MRAM	-Current flow according to magnetization layer direction (electron spin direction)	-Inter-cell interference reduction -High capacity -Same equipment as Dynamic Random Access Memory (DRAM) -Improved price competitiveness, speed and durability
PRAM	-Phase change between crystal and amorphous	-Fine line width -Slower than DRAM but faster than NAND -Slow phase change speed, overcoming power consumption
ReRAM	-Presence or absence of current flow due to change in resistance of insulator	-Insufficient resistance change material optimization

Table 2. Memory properties

	SRAM	DRAM	PCM	ReRAM	MRAM
Latency	300ps ~ ns	10~30ns	~50ns	<10ns	few ns or faster
Endurance	>10^16	>10^15	10^8~10^12	10^6~10^12	>10^15
Retention	volatile	volatile	>10years	>10years	>10years

2. Memory technology

This section details the three representative memory technologies in their characteristics. As shown in Figure 1, each memory technology can be represented by cost and speed. As is well known, SRAM provides high speed and high cost, while DRAM provides relatively slow speed and low cost. MRAM offers similar characteristics as DRAM but with much longer endurance. PRAM and ReRAM put more weight on the cost in the trade-off of speed and cost.

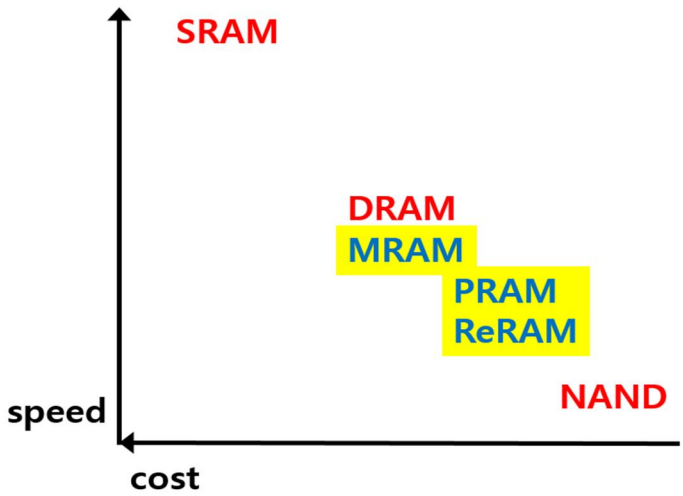

Figure 1. Positioning in memory technology

- Spin Transfer Torque-Magnetic Random Access Memory (STT-MRAM)

Initially, the development was carried out in the form of MRAM, but evolved into STT-MRAM, which simplified the structure of cells to solve the inter-cell interference problem.

- STT-MRAM will be applied as next-generation semiconductor product

MRAM generates magnetic junction tunnel (MTJ) based on the spin property of electrons to realize the flow and interruption of current.

- MRAM is composed of a separate line (Bypass line) 13 to store data, which makes it difficult to implement the density.

In addition, there are problems such as interference between cells and difficulty in maintaining uniformity of insulating film caused by reducing each cell. STT-MRAM stores data by using a magnetization layer through which electrons can pass according to the spin direction of electrons.

- If the magnetization directions are the same (when the spin direction of the electrons is the same), the electrons can pass through and current flows.

- If the magnetization direction is different (if the spin direction of the electron is different), the electron cannot pass through and no current flows.

- STT-MRAM consists of a fixed junction (fixed magnetization direction), a barrier layer (insulation layer) and a free junction (MTJ: Magnetic Tunnel Junction).

STT-MRAM saves data using self-junction tunnel to reduce inter-cell interference and to realize high capacity. It is expected to be a replacement memory for DRAM using the same process as DRAM. However, there is no current price competitiveness, and the problem of speed and durability should be complemented.

- Phase-change Random Access Memory

PRAM is a method of storing data by using phase change between polycrystalline and amorphous. Data is stored when changing from amorphous to crystalline. PRAM is slower than DRAM and faster than Nand Flash, so it is located in the area between DRAM and Nand Flash in terms of speed. PRAM is expected to realize fine line width, but the phase change is slow and power consumption is large.

- Resistive Random Access Memory

ReRAM is a path through which current flows when a high voltage is applied to a large resistor. Data is stored using the characteristic that a filament is formed and the resistance is changed to a state of a conductor. Once a filament is created, it is possible to control the generated passage through voltage and to remove and regenerate the passage. The resistance change material is not currently optimized. ReRAM can be changed to Set (Low Resistance) state by applying Reset (High Resistance) state and 16 (Forward 16) voltage by applying reverse voltage.

3. Memory technology in AI

Current AI systems are designed and developed with system chips that resemble GPUs [7] or Tensorflow Processing Units (TPU) [8] as shown in Figure 2. Figure 2 shows that TPU is designed to provide wide data bandwidth, with memory taking up 29% of the chip size. This memory buffer consumes power in proportion to its size.

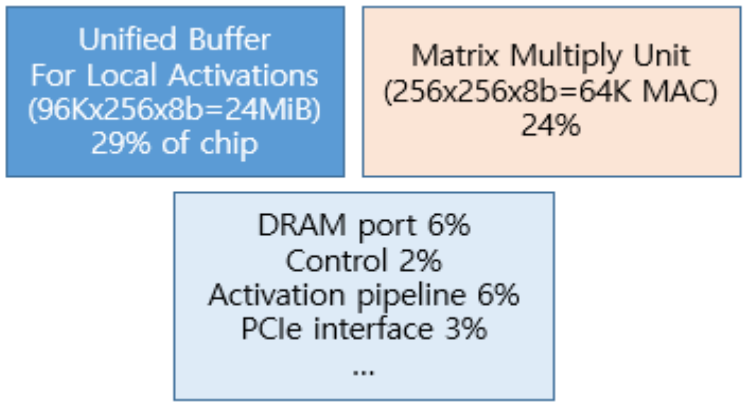

Figure 2. Google's Tensorflow Processing Unit

The total bandwidth of the memory associated with a large processing core is currently exceeding 256GB / s. The need for high-performance, wide memory bandwidth is due to the need for AI to learn a long time using a large amount of data. This is a very important factor for AI because learning time increases exponentially with bandwidth.

AI memory applications can be classified into three stages: mobile devices, general purpose personal computers, and data centers. Each step changes the priority of the system's memory capacity, energy efficiency, and performance. The importance varies in the order of energy efficiency, performance, capacity for mobile devices, performance, energy efficiency, capacity for personal computers, and capacity, energy efficiency, and performance for data centers. One common important fact is that energy efficiency is important in any application. As a result, the required memory technology must be selected and designed to take into account the power efficiency of memory, which accounts for at least 30% of energy efficiency.

4. Conclusion

In this paper, we examined the next generation memory technologies to confirm the memory technologies required for AI applications. As discussed above, AI applications place great emphasis on data bandwidth and energy efficiency, so it is necessary to design optimized memory systems to support them. To achieve this requirement, existing MRAM can replace DRAM and PRAM or ReRAM can replace NAND flash.

Currently, artificial intelligence algorithms are implemented by learning a certain pattern by inputting a large amount of data. As an AI system continue to learn, it can increase accuracy by reducing errors when existing prediction patterns or functions cause errors. This set of learning and reflections is called forward-propagation and backward propagation. A large amount of data streams are generated during the learning process, but data streams are also generated in backward propagation to correct errors. Considering both bidirectional data streams, proper memory development considering energy efficiency, performance, cost and endurance is important. Optimally implemented memory and AI systems could contribute to a very interesting industry such as subway passenger analysis [9], drone applications [10] and etc.

Acknowledgment

This work was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education (NRF - 2018R1D1A1 B07050054).

References

[1] Russell, S., Dewey, D. and Tegmark, M, "Research Priorities for Robust and Beneficial Artificial Intelligence," *AI Magazine*, 36(4), pp. 105-114, 2015.
DOI: https://doi.org/10.1609/aimag.v36i4.2577.

[2] Peter Desnoyers, "What systems researchers need to know about NAND flash," In *Proceedings of the 5th USENIX conference on Hot Topics in Storage and File Systems (HotStorage)*, pp. 1-6, 2013.

[3] Intel Xpoint memory Optane, *www.intel.com/Optane*

[4] Phasse change random-access memory, *https://www.ibm.com/blogs/research/2018/07/phase-change-memory/*

[5] Resistive random-access Memory, *https://www.crossbar-inc.com/technology/reram-overview/*

[6] Magnetoresistive random access memory, *https://searchstorage.techtarget.com/definition/MRAM*

[7] J. D. Owens, M. Houston, D. Luebke, S. Green, J. E. Stone and J. C. Phillips, "GPU Computing," in *Proceedings of the IEEE*, vol. 96, no. 5, pp. 879-899, May 2008.
DOI: 10.1109/JPROC.2008.917757.

[8] N. P. Jouppi et al., "In-datacenter performance analysis of a tensor processing unit," *2017 ACM/IEEE 44th Annual International Symposium on Computer Architecture (ISCA)*, Toronto, ON, pp. 1-12, 2017.
DOI: 10.1145/3079856.3080246.

[9] M.K. Kim, "A Data Design for Increasing the Usability of Subway Public Data," *International Journal of Internet, Broadcasting and Communication (IJIBC)*, Vol.11, No.4, pp. 18-25, 2019.
DOI: http://dx.doi.org/10.7236/IJIBC.2019.11.4.18.

[10] D. Cho, "A Study on Efficient Use of Dual Data Memory Banks in Flight Control Computers," *International Journal of Internet, Broadcasting and Communication (IJIBC)*, Vol.9, No.1, pp. 29-34, 2017.
DOI: https://doi.org/10.7236/IJIBC.2017.9.1.29.