SEMICONDUCTOR MEMORIES FOR LOW-POWER AND LOW-COMPUTE HARDWARE-ORIENTED ARTIFICIAL INTELLIGENCE SYSTEM

SOOMIN KIM[1], YEJI LEE[1], SO WON SON[1], SUNGJUN KIM[2], SOO YOUN KIM[3], MYOUNGGON KANG[4], AND SEONGJAE CHO[1,*]

[1]Division of Electronic and Semiconductor Engineering, Ewha Womans University, Seoul 03760, Republic of Korea
[2]Division of Electronics and Electrical Engineering, Dongguk University, Seoul 04620, Republic of Korea
[3]Department of System Semiconductor, Dongguk University, Seoul 04620, Republic of Korea
[4]Department of Intelligent Semiconductor Engineering, University of Seoul, Seoul 02504, Republic of Korea
[*]E-MAIL: felixcho@ewha.ac.kr

Abstract:

The development of artificial intelligence (AI) technologies for reasoning based on big data is rapidly advancing day by day. Moving beyond large language models (LLMs), recent technological trends show the emergence of large vision models (LVMs), indicating that the application scope of AI is expanding at an accelerated pace. Currently, most AI services are implemented through software technologies. However, from the perspective of energy saving and environmental pollution, it is a crucial turning point where a shift to hardware-oriented AI technology must take place. Hardware-oriented AI aims to move away from the conventional series high-speed operation, towards low-power computing technologies that maximize computational concurrency. In order to achieve this goal, changes in computing architecture are necessary, with semiconductor memory technology playing a central role. Simultaneously, recent research indicates that high-speed, large-scale computation systems naturally lead to increased system temperatures, which can produce gases harmful to human health. Although these goals differ in terms of the original starting points, all these technological objectives share a common aim of low-power and small-number computing. This paper examines next-generation AI computing technologies based on large-capacity memory technologies, specifically dynamic random-access memory (DRAM) and flash memories built on relatively mature Si fabrication processing suitable for chip production, evaluating pattern recognition accuracies.

Keywords:

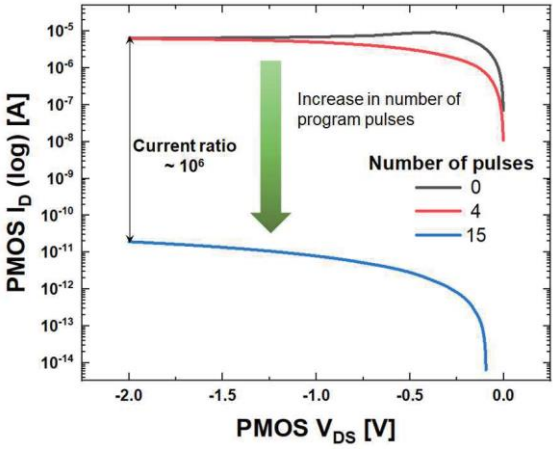
Artificial intelligence (AI); Energy saving; Hardware-oriented AI; Tracking; Estimation; Computational concurrency; Si processing; Global energy and environment

1. Introduction

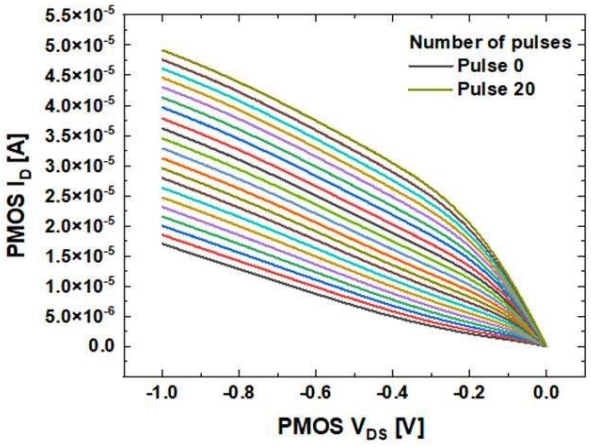
The advancement of semiconductor devices is driven by the quest for faster, more efficient computers that enhance everyday life. This progress has largely depended on logic devices shrinking through technology nodes, enabling high-speed and low-power operations [1]. While lighter, faster processors remain paramount, the exponential growth of data demands innovative solutions beyond traditional serial communication and processing [2]. Challenges such as inherent logic switching delays [3], resistance-capacitance (RC) delay in metallic interconnects [4], and communication bottlenecks between logic and memory limit performance [5]. Controlling the first two is constrained by physical and fabrication limits, but reducing communication latency remains possible through novel devices and architectures that maximize computing concurrency. Consequently, semiconductor memories, once secondary to processor developments, are now moving to the center of future computer architectures [6]. It is anticipated that the hardware-oriented artificial intelligence (AI) will play a crucial role in alleviating environmental stressors, thereby supporting the dignity of human life. This paper explores the pivotal roles and requirements of semiconductor memories in the hardware-oriented AIs in neuromorphic systems and processing-in-memory (PIM).

2. Hardware-Oriented AI with Si Memories

While various approaches exist, implementing neurons as integrated circuits and synapses with high-density memory is the most practical combination for the physical realization of hardware AI [7]. It is because of the fact that neurons require nonlinear signal processing and high current-handling capabilities, while synapses must store a huge number of parameters, making high-density memory advantageous. Also, to build AI systems in chips, highly matured Si technology is essential. Therefore, hardware-oriented AI is intrinsically linked to Si memories, with dynamic random-access memory (DRAM) and flash memory serving as prominent options.



(a)



(b)

FIGURE 1. 2T DRAM operation characteristics. (a) Memory operation mode (binary) and (b) synaptic device operation mode with multiple weights.

2.1. Hardware-Oriented AI with DRAM Technology

Previously, neuromorphic chips with fully compatible with Si CMOS processing were primarily implemented by static random-access memory (SRAM) for synapses. While Ai has predominantly advanced through software, machine learning—especially deep neural networks (DNNs)—requires large datasets. To succeed as hardware neuromorphic systems, synaptic devices must scale to high-density arrays. However, bulky SRAMs, composed of six transistors, are impractical for this purpose [8,9], limiting applications [10]. In contrast, DRAM, a volatile memory with higher cell scalability, has recently gained attention as a potential synaptic device. Studies with recency show suitability of DRAM for accelerators in convolutional neural networks (CNN) and recurrent neural networks (RNN) due to its area and cost efficiencies [11]. Notably, even in CNN architecture using

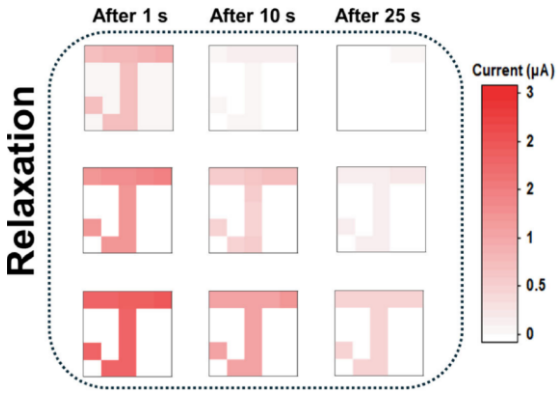


FIGURE 2. Pattern recognition capability over time of an IGZO 2T DRAM.

DRAM, the memory is primarily used for storing compressed feature maps and kernels, not for synaptic computation [11]. Until recent dates, DRAM has not been widely adopted in neuromorphic computing largely due to its requirement for periodic refresh operations and destructive read (inference) operations. Since neuromorphic architectures prioritize energy efficiency and massive parallelism for large-scale data processing, the time and bandwidth loss associated with DRAM refresh can pose serious limitations. Recently, a novel DRAM cell featuring two independent MOSFETs without a capacitor has been developed [12,13]. One MOSFET (write transistor) manages learning functions (potentiation and depression), while the other one (read transistor) independently handles inference, enabling non-destructive operation and significantly improved data retention. Fig. 1(a) shows the binary operation with a properly high operation voltage on the two-transistor (2T) DRAM cell. There is no intermediate state between two read current levels. On the other hand, as can be confirmed by Fig. 1(b), 2T DRAM cell can perform synaptic functions of learning with 16-level (4-bit) resolutions or electrical weights [12]. This dual-mode operation might allow the cell to function either as a standard DRAM or as a synaptic device for neuromorphic applications, depending on the programming voltage. One weak point of a DRAM cell is there is much room to improve data retention. With the genuinely existing issue of retention loss, the memory bandwidth is shrunken. This can be also a technological limit in neuromorphic or processing-in-memory (PIM) applications since the learning (program and erase operations) and inference (read operation) events are required to take place more frequently in the data-intensive operations in the AI chip. Thus, the great deal of effort has been dedicated to elongation of retention time. It was recently reported that introducing indium-gallium-zinc-oxide (IGZO) as the channel material of the write transistor and a long retention time of 25 in pattern recognition was reached (Fig. 2) [14].

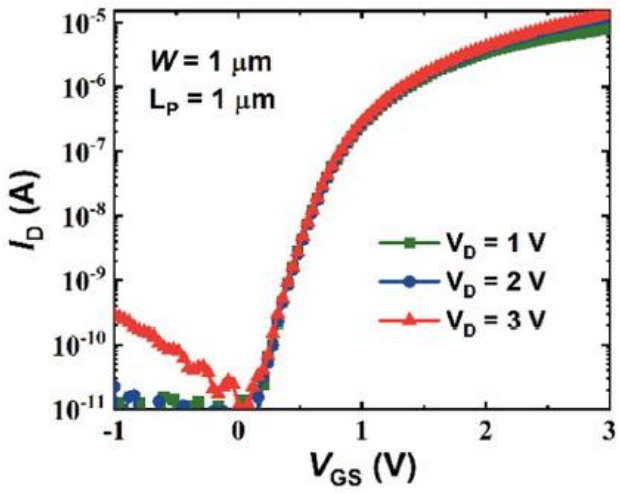


FIGURE 3. Transfer characteristics of a CTF synaptic device with poly-Si channel. The on-state current is in the order of microamperes per unit width.

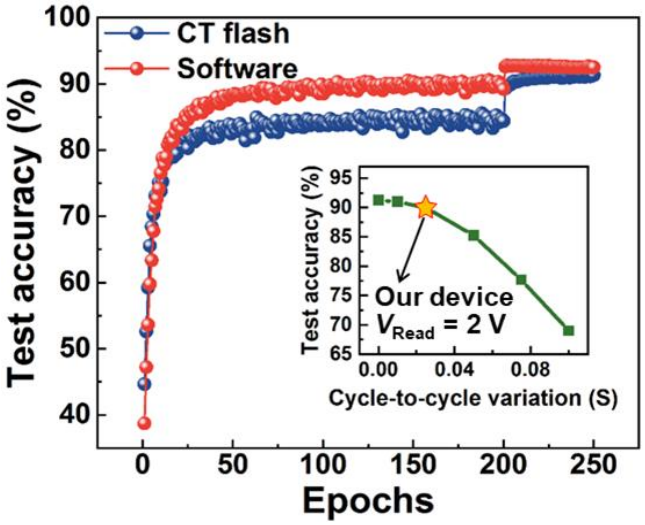


FIGURE 4. Test accuracy as a function of number of epochs. (inset: accuracy drop caused by cycle-to-cycle weight variation of a CTF synapse)

2.2. Low-Power Inference in Charge-Trap Flash (CTF)

Although SRAM and DRAM currently serve as weight storage media, essentially functioning as synapses, for hardware-oriented AI inference operations, the number of synapses corresponds to the number of AI parameters, making it advantageous to implement the synapses using high-density memory technologies. From this perspective, among the commercially available Si-based memory technologies, charge-trap flash (CTF) is the most strategic for implementation of hardware AI chip [15]. However, its rather slow program and erase operations realized by Fowler-Nordheim (FN) tunneling, the practical memory bandwidth is reduced and the maximal operation concurrency is substantially threatened. Also, the high operation voltages of CTF calls for the complicated peripheral circuits. As a result, speed and area efficiencies need to be improved for flash technology to actively come to presence in realizing the hardware-oriented AI chip. The CTF cell is typically fabricated with a structure that includes a poly-Si channel. The primary advantage of using poly-Si instead of crystalline Si for the channel is the strong potential for low-power operation. As shown in Fig. 3, CTF synaptic devices fabricated in a previous study conduct currents on the order of several microamperes per unit channel width [16]. Considering that the International Roadmap for Devices and Systems (IRDS) suggests that a transistor with a current of 1 mA/μm could serve as a threshold between high-performance (HP) and low-power (LP) classifications [17], it is explicitly revealed that the current levels observed in the fabricated CTF cell fall within the low-power electronic device domain.

Fig. 4 presents the results of simulations performed at a higher level using key operational parameters extracted at the cell level from the fabricated CTF synaptic device [16]. The simulation evaluates the pattern recognition accuracy of a presumable AI chip when an array composed of the fabricated CTF synapses is integrated into it. Fig. 4 shows the test accuracy as a function of number of epochs and it is encouraging to observe that the accuracy reaches about 85% (5% below compared with the purely software-based test) at a number of epochs above 50. The inset depicts test accuracy as a function of cycle-to-cycle variation in synaptic weight (electrical conductance) of the fabricated CTF cell. Although cell-to-cell conductance variation in flash memory based on Si processing is relatively smaller compared with other emerging nonvolatile memories, a deliberately large variation was assumed in the simulation works above. As one might expect, if the variation in the targeted weight values during repeated synaptic device operations is significant, the pattern recognition accuracy shows a rapid monotonic decrease. While minimizing cell-to-cell and cycle-to-cycle variations is fundamentally important, it becomes even more critical in hardware-oriented AI applications than in the conventional electronic systems. If the inference voltage can be increased, thereby raising the level of inference current, the relative magnitude of the variation decreases. Therefore, using a sufficiently high inference voltage would be beneficial. From this perspective, it is revealed by the inset that inference energy efficiency and learning accuracy are in a trade-off relation. Thus, from a system-level operation standpoint, it would be desirable to devise a smart operation scheme in parallel that balances these aspects effectively.

3. Low-Compute Hardware AI for Environment

As semiconductor devices continue to be miniaturized to achieve higher computational speeds, concerns have emerged regarding the environmental impact of such advanced electronics. One particularly alarming issue is the potential release of harmful gases from computers operating high-speed, ultra-scaled AI hardware based on the highly scaled transistors in the conventional architecture. These emissions are not generated by the transistors themselves during normal operation, but rather as a consequence of excessive heat buildup, material degradation, and outgassing from components such as printed circuit boards (PCBs), thermal interface materials, and packaging polymers. When devices operate under extreme thermal and electrical stress (conditions common in densely packed AI accelerators) certain volatile organic compounds (VOCs) or fluorinated gases used in manufacturing or thermal control can be released into the surrounding environment. This highlights an often-overlooked environmental cost of sustaining ever-faster, high-power computational loads [18]. In light of this, next-generation computing technologies stated in the previous chapters and those viewed from an environmental perspective may have different motivations, but they share the same overarching goal. Alternative computing paradigms such as neuromorphic computing and PIM offer a more sustainable path forward. These architectures fundamentally deviate from the traditional von Neumann architecture, which is heavily reliant on sequential processing and frequent data shuttling between processor and memory domains – a process that contributed significantly to energy consumption. Instead, hardware-oriented AI systems in the neuromorphic and PIM approaches prioritize parallelism and local computation, enabling a large number of low-power operations to be performed simultaneously and reducing the need for constant data traffic. Rather than maximizing the number of operations per second at all costs, shifting towards low-compute and extremely parallel architectures not only align better with the energy constraints of real-world AI deployment but also mitigates environmental impact. As such, contemplating the hardware-oriented AI from an ecological perspective may be more crucial than advancing its computational capabilities.

4. Conclusion

In this study, the limitations of traditional von Neumann architectures, particularly RC delays, logic switching latency, and memory bottlenecks, are addressed through emerging hardware-oriented AI approaches such as neuromorphic computing and PIM. These paradigms leverage Si-based memories, DRAM and flash, not just for storage but also for active involvement into actual computation. Novel 2T DRAM designs with IGZO channels show promising results, achieving up to 25 seconds of data retention and enabling 4-bit synaptic weight resolution. CTF memory, offering ultra-low-power inference operation (<1 μA/μm), achieves ~85% pattern recognition accuracy after 50 training epochs, though performance is sensitive to synaptic variation. Environmentally, these architectures present a sustainable alternative to heat-intensive and highly-scaled processors, which risk releasing harmful gases. By embracing parallel and low-compute designs, hardware-oriented AI chip meets both technological and ecological goals. It is highlighted that The future of AI lies not in maximizing raw speed but in designing smarter and energy-conscious architecture that harmonize performance with planetary human health.

Acknowledgements

This work was supported by National R&D Programs through the National Research Foundation of Korea (NRF) grant funded by the Ministry of Science and ICT under RS-2023-00258527, RS-2024-00402495, RS-2024-00356939, and 2022M3I8A1077243.

References

[1] Q. Zhang, H. Deng, and K. Song, “Latest VLSI Techniques for 3nm Technology for Building Efficient AI Chips,” *An International Journal of Fusion of Multidisciplinary Research*, vol. 5, no. 2, pp. 654–670, Sep. 2024.

[2] A. Jaiswal, I. Chakraborty, A. Agrawal, and K. Roy, “8T SRAM Cell as a Multibit Dot-Product Engine for Beyond Von Neumann Computing,” *IEEE Trans. Very Large Scale Integr. VLSI Syst.*, vol. 27, no. 11, pp. 2556–2567, Aug. 2019.

[3] A. Khakifirooz and D. A. Antoniadis, “MOSFET Performance Scaling – Part II: Future Directions,” *IEEE Trans. Electron Devices*, vol. 55, no. 6, pp. 1401–1408, Jun. 2008.

[4] M. Darmi, L. Cherif, J. Benallal, R. Elgouri, and N. Hmina, “Integrated Circuit Conception: A Wire Optimization Technic Reducing Interconnection Delay in Advanced Technology Nodes,” *Electron.*, vol. 6, no. 4, pp. 78-1–78-11, Oct. 2017.

[5] P. Montuschi, Y.-H. Chang, and V. Piuri, “In-Memory Computing: The Emerging Computing Topic in the Post-von Neumann Era,” *Comput.*, vol. 56, no. 10, pp. 4–6, Oct. 2023.

[6] S. Cho, “Volatile and Nonvolatile Memory Devices for Neuromorphic and Processing-in-Memory

Applications," *J. Semicond. Technol. Sci*., vol. 22, no. 11, pp. 30–46, Feb. 2022.

[7] B. Jeong, J. Lee, S. Lee, S. Lee, Y. Son, and S. Y. Kim, "A 240 FPS In-Column Binarized Neural Network Processing in CMOS Image Sensors," *IEEE Trans. Circuits Syst. II Express Briefs*, vol. 70, no. 10, pp. 3907–3911, Jul. 2023.

[8] F. Akopyan, *et al*., "TrueNorth: Design and Tool Flow of a 65 mW 1 Million Neuron Programmable Neurosynaptic Chip," *IEEE Trans. Comput. Aided Des. Integr. Circuits Syst*., vol. 34, no. 10, pp. 1537–1557, Oct. 2015.

[9] M. Davies, A. Wild, G. Orchard, Y. Sandamirskaya, G. A. F. Guerra, P. Joshi, P. Plank, and S. R. Risbud, "Advancing Neuromorphic Computing With Loihi: A Survey of Results and Outlook," *Proc. IEEE*, vol. 109, no. 5, pp. 911–934, May 2021.

[10] A. G. Andreou, *et al*., "Real-time sensory information processing using the TrueNorth Neurosynaptic System," *Proc. 2016 IEEE International Symposium on Circuits and Systems (ISCAS)*, Montreal, QC, Canada, pp. 22–25, May 2016.

[11] T. Delbruck and S.-C. Liu, "Data-Driven Neuromorphic DRAM-based CNN and RNN Accelerators," *Proc. 2009 Sig. Proc. Soc. Asilomar Conference on Signals, Systems, and Computers*, pp. 1–7, Asiloma, CA, USA, Nov. 3–6, 2019.

[12] S. Baek, B. E. Yoo, I. Lee, and S. Cho, "Design of Compact 2T(0C) DRAM Cell Allowing Nondestructive Read Operation and Glance at Its Applications as Synaptic Device," *Proc. 2021 IEIE Summer Conf.*, pp. 515–516, Jeju, Korea, Jun. 30–Jul. 2, 2021.

[13] S. Cho and S. Baek, "Two-Transistor Memory Cell, Synaptic Cell and Neuron Mimic Cell Using the Same and Operation Method Thereof," *Korean Patent*, 10-2766751, Feb. 2, 2025.

[14] J. Jang, S. Kim, S. Park, S. Kim, S. Kim, and S. Cho, "Leaky 2T Dynamic Random-Access Memory Devices Based on Nanometer-Thick Indium-Gallium-Zinc-Oxide Films for Reservoir Computing," *ACS Appl. Nano Mater*., vol. 7, no. 19, pp. 22430–22435, Oct. 2024.

[15] Md. H. R. Ansari, U. Mohanan Kannan, and S. Cho, "Core-Shell Dual-Gate Nanowire Charge-Trap Memory for Synaptic Operations for Neuromorphic Applications, *Nanomater*., vol. 11, no. 7, pp. 1773-1–1773-14, Jul. 2021.

[16] M.-K. Park, J. Hwang, S. Kim, W. Shin, W. Shim, J.-H. Bae, J.-H. Lee, and S. Cho, "Charge-trap synaptic device with polycrystalline silicon channel for low power in-memory computing," Sci. Rep., vol. 14, pp. 29089-1–29089-14, Nov. 2024.

[17] IEEE International Roadmap for Devices and Systems (IRDS™), online available at https://irds.ieee.org/editions.

[18] A. Tabbkh, L. A. Amin, M. Islam, G. M. I. Mahmud, I. K. Chowdhury, and Md. S. H. Mukta, "Towards sustainable AI: a comprehensive framework for Green AI," *Discover Sustainability*, vol. 5, pp. 408-1–408-14, Nov. 2024.