JOURNAL OF SEMICONDUCTOR TECHNOLOGY AND SCIENCE, VOL. 22, NO. 1, FEBRUARY, 2022
https://doi.org/10.5573/JSTS.2022.22.1.30

ISSN(Print) 1598-1657
ISSN(Online) 2233-4866

Volatile and Nonvolatile Memory Devices for Neuromorphic and Processing-in-memory Applications

Seongjae Cho*

Abstract—**The motivation for driving semiconductor devices can be found in the development of advanced computers which can contribute to the betterment in our daily lives. The contribution has been largely made by semiconductor logic devices traveling the pavements identified as technology nodes for device shrinkage that enables high-speed and low-power operations. Lighter and faster processors are the everlasting goals in electronics and computer science, and have been concerned with logic technologies. However, the vast amount of data that should be dealt are consistently requiring an innovative way out of the conventional serial data communication and processing. Data need to be processed in a shorter time but the irreducibilities in logic switching time, data propagation time in metallic interconnection accompanying RC delay, and the time amount spent in the serial communication between logic and memory units should be quenched. It is quite hard to control the former two factors which are largely determined by physical limits and fabrication technology ones in recent days but the latter still has room for reduction by novel devices and architectures specifically designed for maximizing the parallelism in data processing and communication. The semiconductor memories let aside the advancements in processor technologies now is being moved to the center of renovation toward the future computers in the ultimate architecture. In this review, the roles and requirements of semiconductor memories for memory-oriented processors are investigated in the highlights of applications in the neuromorphic system and processing-in-memory (PIM) architectures.**

Index Terms—**Semiconductor devices, semiconductor memories, data processing, computer architecture, neuromorphic system, processing-in-memory (PIM), memory processing unit (MemPU)**

I. INTRODUCTION

The actual effective speed of a computer system is determined by speed of memory, and further, that of communication between processing and memory units. It is an undoubted fact that the intrinsic gate delay governs system speed most fundamentally, but we are not living in the era in which the processing speed of a central processing unit (CPU) is determined by the speed of transistor switching although the great deal of effort has been dedicated to shrinkage of transistor for higher switching speed and low power consumption. For being capable of accommodating the gigantic amount of data, stronger parallelism has been consistently required. Parallel computing, high-performance computing (HPC), distributed computing, and grid computing can be thought as the effort for increasing the system speed by physical segmentations of computers over space, operating in the time-division manner [1-4], which have been prevalent. In recent times, such computers are shrinkun into a chip with the highly scaled parallelism, which can be easily found in the contemporary multi-core CPUs and many-core graphic processing units (GPUs) [5-8]. However, these technologies are highly

Manuscript received Dec. 15, 2021; reviewed Feb. 20, 2022; accepted Feb. 21, 2022

Department of Electronic Engineering, Gachon University, Sujeong-gu, Seongnam-si, Gyeonggi-do 13120, Korea
E-mail : felixcho@gachon.ac.kr

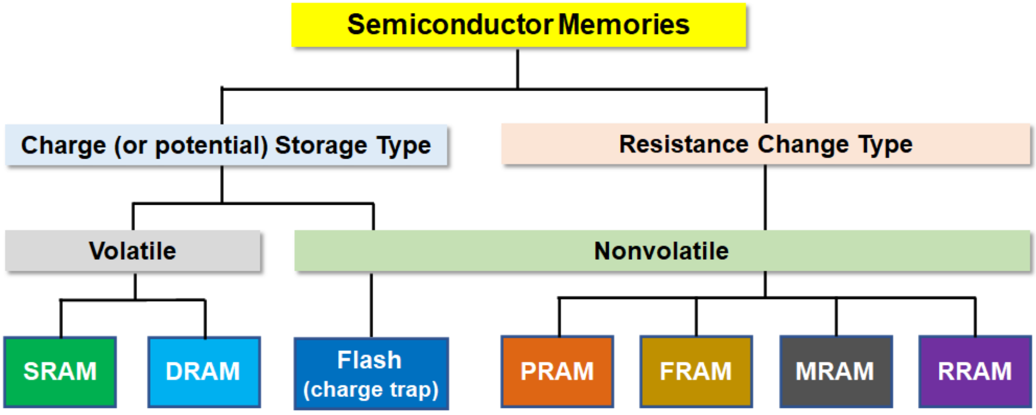


Fig. 1. Hierarchy in semiconductor memories for the neuromorphic and PIM applications.

dependent on scaling technology of transistors and what crucially matters is the logic operation speed, not taking the actual system speed determined at the level of massive nonvolatile memories into the serious consideration. The computing performances have been referred as the result of semiconductor logic technology but the importance of memory technologies is getting larger and larger as the high-performance computers and hardware-driven artificial intelligence (AI) become more big-data-driven and require expedited communication between the processor core and the ultra-high-density memory area [9]. In this review, volatile and nonvolatile memory devices making up the most fundamental functional cells in the advanced computer architectures are surveyed in the highlights of their applications in the hardware-driven neuromorphic systems and processing-in-memory (PIM).

Semiconductor memories can be categorized by two criteria as schematically shown in Fig. 1: (i) whether it is charge-storage type or resistance-chancing type and (ii) whether it is volatile or nonvoltaile. Great majority of Si memory devices are found in the charge (or potential) storage type including static random-access memory (SRAM), dynamic random-access memory (DRAM), and flash memory. Although floating-gate (FG) structure had the great majority in the past flash memory technologies and it can be still found in the microcontroller units (MCUs) embedding the FG flash memories owing to its perfect Si processing compatibility, the predominence is taken by the charge-trap flash (CTF) technology in recent times. In the charge storage memory regime, the device evolutions have been progressed with a relatively high emphasis on novel device stucturing since the base materials that can be accommodated in the fabrication facility for mass chip production are not unlimited and the Si processing technologies are highly matured. On the other hand, in the regime of resistance change memories, the mateirals are being sought without ceasing and the development and optimization of process architecture are of parallel concern. It needs to be clarified that resistance change type and resistive-switching random access memory (RRAM) do not have the same definition but they have different set and subset relations as clearly grasped by Fig. 1. The resistive change memory refers to all the memories in which the state changes can be altered by the change in resistance, or equivalently, that in conductance. Phase-change random-access memory (PRAM), ferroelectric RAM (FRAM), magnetic RAM (MRAM), and resistive-switching RAM (RRAM) belong to resistance change memory technology.

The following sections have been organized covering all the hiararchies: neuromorphic applications based on charge storage volatile memories, SRAM and DRAM, are surveyed in Chapter II. Those with nonvolatile memories are investigated in Chapter III, which is more specifically divided into Chapter III. 1 for charge storage CTF and Chapter III. 2 for resistance change memories including all of PRAM, FRAM, MRAM, and RRAM, respectively.

II. VOLATILE MEMORY CELLS FOR NEUROMORPHIC APPLICATIONS

Neuromorphic computing is a new way of computing mimicking the behaviors of nervous system. The most

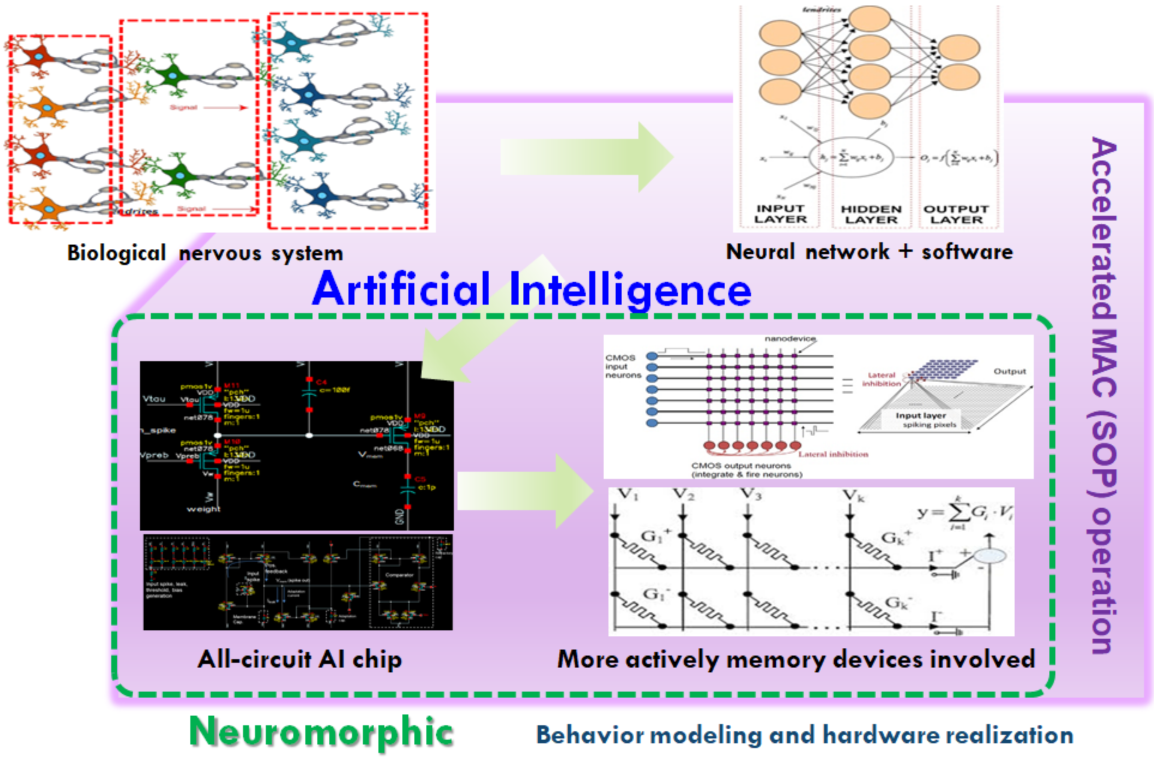

Biological nervous system
Neural network + software
INPUT LAYER
HIDDEN LAYER
OUTPUT LAYER
Artificial Intelligence
Accelerated MAC (SOP) operation
All-circuit AI chip
More actively memory devices involved
$y=\sum G_i \cdot V_i$
Neuromorphic
Behavior modeling and hardware realization

Fig. 2. Artificial intelligence and neuromorphic computing.

fundamental nervous behavior is broken down into the multiplication-and-accumulation (MAC) operations that take place between neurons as schematically shown in the upper part in Fig. 2. Neuromorphic computing is a forward step by which AI can be implemented in a more physical way so that the MAC operations can be carried out with higher volume and energy efficiencies [10]. For this goal, more specifically designed hardwares - integrated circuits and devices - are necessitated as shown in the lower part in Fig. 2. The early AI was realized in the highly algorithm-intensive manner, in which the volume and energy efficiencies were not substantially considered [11]. More hardware-oriented state-of-the-art neuromorphic chips have been incessantly released with the full Si CMOS processing compatibility [12-14], where the synapses were made of static random-access memories (SRAMs). AI has been primarily led by software and is pursuing machine learning as can be schematically shown in Fig. 3. Deep neural network (DNN) is a widely admitted way to realize machine leaning that essentially requires big data. Thus, to be a successful hardware neuromorphic system, the synaptic device or cell needs to equip higher scalability toward a high-density synapse array. However, the bulky SRAM composed of 6 transistors is not strategic to practically achieve the goal [13, 14], and as the result, applications can be quite limited [15]. Although it has been rarely reported until recent date in comparison with SRAM, dynamic random-access memory (DRAM) is another volatile meory cell that can be also utilized for the hardware-driven neuromorphic system as the synapse with higher cell scalability. It was reported that DRAM can be used in the accelerator for either convolutional neural network (CNN) or recurrent neural network (RNN) due to the area and cost effectiveness of DRAM [16]. Even in case of the architecture of a CNN accelerator employing DRAM, the DRAM domain is not used for synaptic computing but for providing the compressed feature maps and kernal as

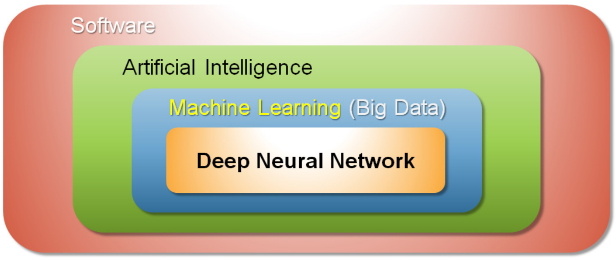


Fig. 3. Background and orientation of the artificial intelligence.

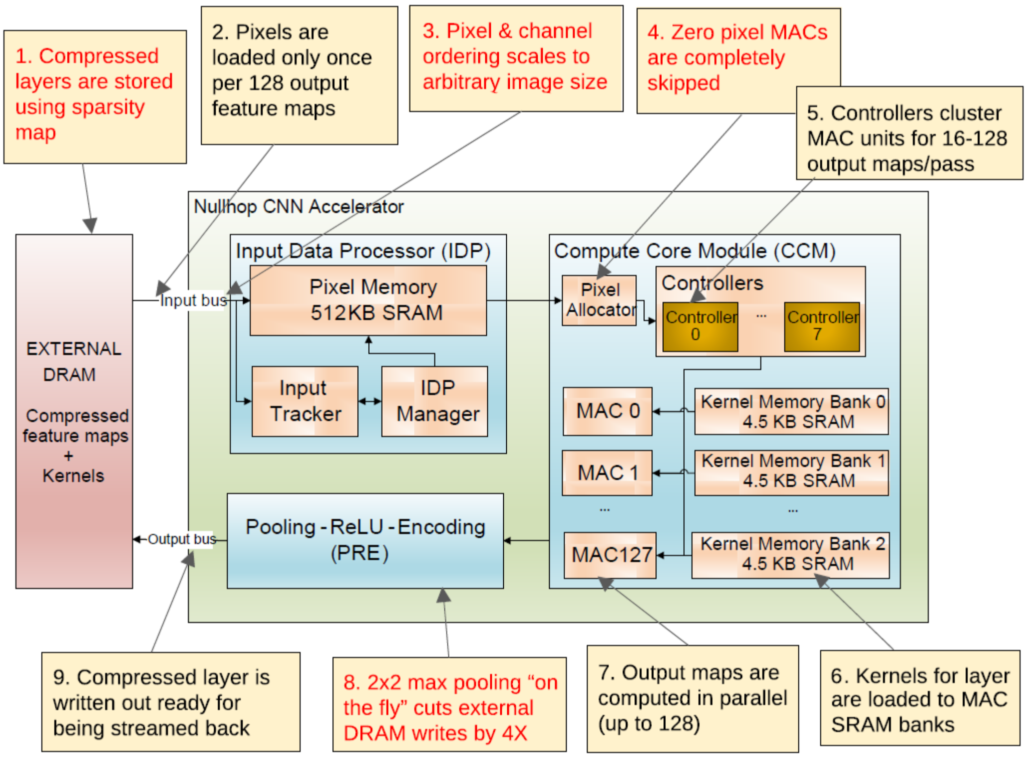

1. Compressed layers are stored using sparsity map
2. Pixels are loaded only once per 128 output feature maps
3. Pixel & channel ordering scales to arbitrary image size
4. Zero pixel MACs are completely skipped
5. Controllers cluster MAC units for 16-128 output maps/pass
Nullhop CNN Accelerator
EXTERNAL DRAM
Compressed feature maps + Kernels
Input bus
Output bus
Input Data Processor (IDP)
Pixel Memory 512KB SRAM
Input Tracker
IDP Manager
Compute Core Module (CCM)
Pixel Allocator
Controllers
Controller 0
...
Controller 7
MAC 0
MAC 1
...
MAC127
Kernel Memory Bank 0 4.5 KB SRAM
Kernel Memory Bank 1 4.5 KB SRAM
...
Kernel Memory Bank 2 4.5 KB SRAM
Pooling - ReLU - Encoding (PRE)
9. Compressed layer is written out ready for being streamed back
8. 2x2 max pooling "on the fly" cuts external DRAM writes by 4X
7. Output maps are computed in parallel (up to 128)
6. Kernels for layer are loaded to MAC SRAM banks

Fig. 4. Architecture of a CNN accelerator with DRAM [16].

schematically shown in Fig. 4 [16]. It has not been explicitly addressed but the reason that DRAM has not been actively adopted for the neuromorphic computing can be found from the fact that the periodic refresh operations are required in the conventional DRAM cell.

Neuromorphic computing architectures are specifically designed for higher energy efficiency and superb parallelism in big data processing. The loss of time and data bandwidth in the DRAM synapse array can be seriously concerned. Thus, if DRAM cells can be adopted in the neuromorphic applications as the synaptic units, the issue of data retainability should be resolved. A novel DRAM cell featuring two independent MOSFET devices, without capacitor, has been recently invented and presented [17, 18]. The first MOSFET takes charge of learning operations (potentiation and depression) and the second one takes charge of inference only, by which non-destructive inference operation and substantially increased data retention are warranted. Further, the invented DRAM cell can be operated in the dual modes: one for the stand-alone DRAM and the other for neuromorphic application depending on the magnitude of voltage pulse for program and erase operations. Fig. 5 demonstrates the output curves of the second MOSFET where the inference operation takes place. The functionality as a synapse cell with plausibly linear weight modulation capability in terms of number of learning pulses is clearly demonstrated in Fig. 5.

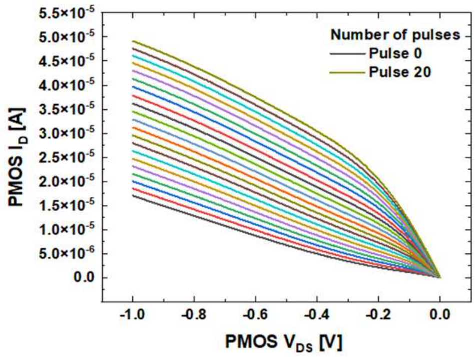


Fig. 5. Synaptic operation of a novel DRAM cell [17].

Although the usefulness and functions of the short-term memory (STM) in the hardware neuromorphic system can be differed from those in the biological nervous system [19, 20], STM is essential in design and realization of time-series neuromorphic system based on RNN [21-23]. Thus, the STM-oriented neuromorphic systems can be surely realized by volatile memories including SRAM and DRAM as surveyed above, and higher data capacity, energy efficiency, the time-invariant weight retainability can be realized by introducing the nonvolatile memory synapses as will be reviewed in the subsequent sections.

III. Nonvolatile Memory Cells for Neuromorphic Applications

1. Charge-trap Memory Synaptic Devices

All-circuit AI chip in Fig. 2 can be categorized into neuromorphic system since area and energy efficiencies are enhanced, in comparison with the software-driven AI, by the approach of more specific hardware design. Since the all-circuit AI chip has the Si processing compatibility, it had a higher chance to reach chip production earlier. However, a functional synaptic unit is composed of plural transistors so that there is much room to increase the area and energy efficiencies. It should be correct to express memory cells when it comes to SRAM or DRAM, the volatile memories, rather than memory devices. However, when dealing with nonvolatile memories, a single device can function as one synapse. In consequence, the nonvolatile memory synapse has higher device scalability and array density. Also, nonvolatile memories are superior to voltaile ones with regard to energy efficiency when they weave the synapse arrays for neuromorphic systems. An early single-device nonvolatile synapse was invented in the structure of floating body with charge-trap layer [24]. The Si-based floating-body synaptic transistor (SFST) is capable of both STM and long-term memory (LTM) functions. SFST can be specifically understood as the combination of one-transistor (1T) DRAM and charge-trap flash (CTF) memory for short- and long-term memories, respectively. Fig. 6 schematically shows the principles for the synaptic operations. The electron-hole pairs are generated by hot-carrier-induced impact ionization. The electrons are drifted into the drain junction and the holes are accumunlated in the floating body. A recent research results show that diffusion has the predominance over drift and recombination in determining retention of data in 1T DRAM [25]. In other words, the accumulated holes in the *p*-type body vanish by extremely fast diffusion of holes into the source and drain junctions, unless the potentiation pulses are repeated with short enough intervals. By this accumulation and fast diffusion, threshold voltage of the SFST is temporarily elevated and comes back to the initial value, which realizes the STM. Repeated potentiation pulses increase the population of the holes accumulated in the floating body and the holes have higher probability to be injected into the charge-trap layer by tunneling. By the trapped holes, the threshold voltage becomes invariant if there is no intended depression (erase) operation. The SFST necessitates a floating body for realizing the STM function but the holes can be also temporarily stored by preparing other type of storage. Fig. 7(a) schematically shows the semi-floating-gate synaptic transistor (SFGST) which can be fabricated on the bulk Si wafers [26]. The potentiation takes place by band-to-band tunneling of holes from the channel into the semi-floating gate (SFG) of which one end is connected to the channel as shown in Fig. 7(b) and (c). STM is realized due to diffusion of the holes out of the SFG to the channel. It would be essential

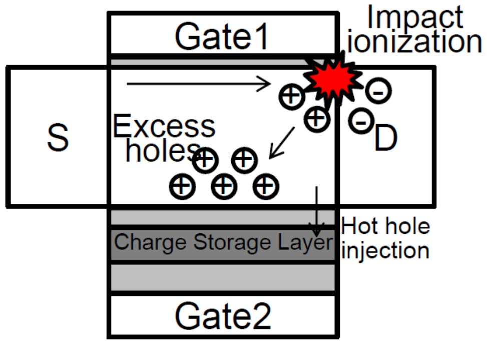


Fig. 6. Si-based floating-body synaptic transistor (SFST) [24].

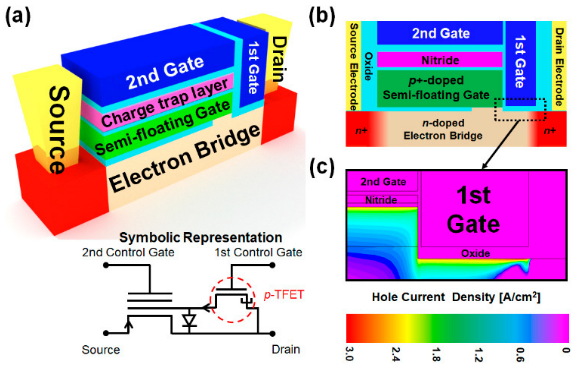


Fig. 7. Device structure and potentiation process of semi-floating-gate synaptic transistor (SFGST): (a) Aerial view of the SFGST and its circuit symbol representation; (b) Cross-sectional view of the device; (c) Contour of hole current density during the potentiation through band-to-band tunneling [26].

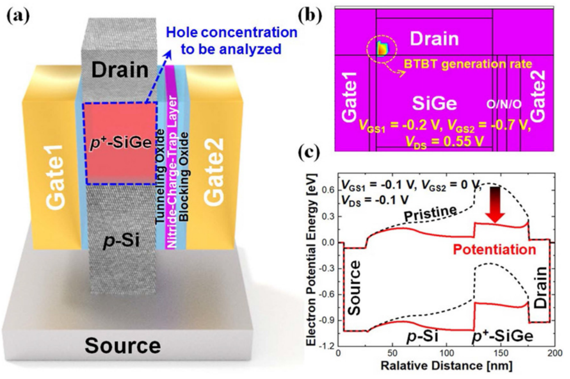


Fig. 8. Quantum-well charge-trap synaptic transistor (CTS): (a) Schematic of the device structure; (b) Tunneling rate in the channel direction investigated by device simulation; (c) change in energy-band diagram during a potentiation [28].

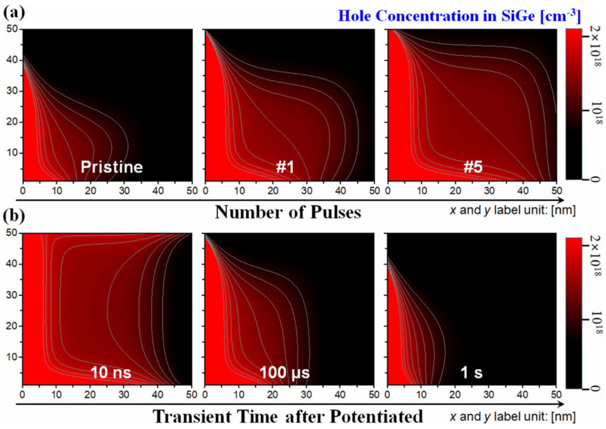


Fig. 9. (a) Short-term memory functionality of CTS. Increase of the hole concentration in the whole SiGe layer with potentiation pulse number; (b) the decay in the absence of a pulse [28].

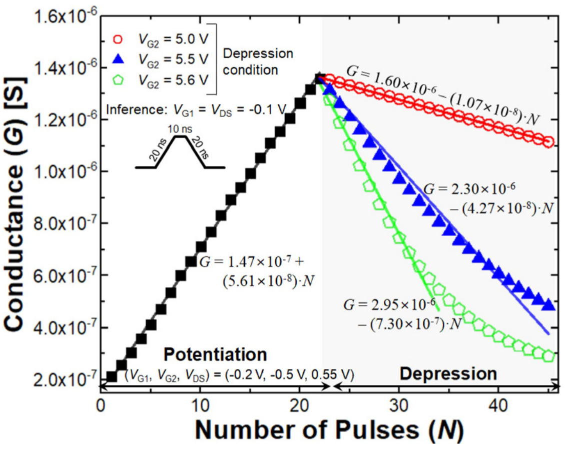


Fig. 10. Highly linear conductance change of the CTS device with regard to number of learning pulses. The earlier 23 pulses are for potentiation and the latter ones are for depression [28].

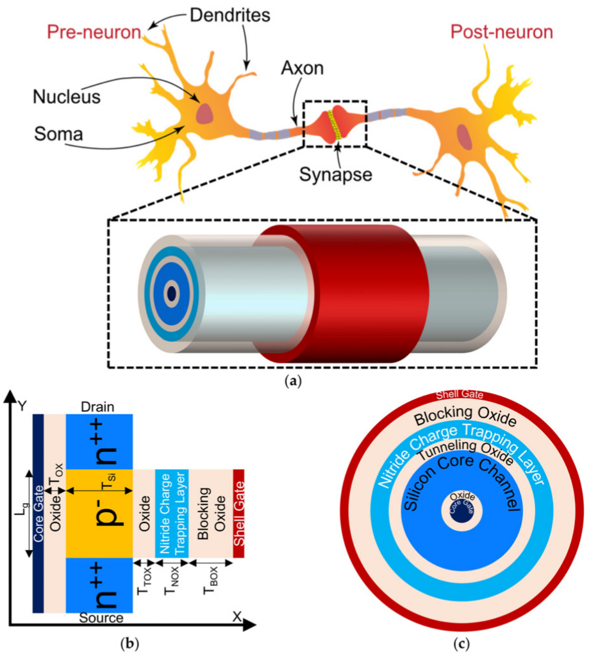


Fig. 11. Device structure of core-shell dual-gate (CSDG) nanowire synaptic transistor: (a) Three-dimensional view. Cross-sectional views; (b) along; (c) across the channel [29].

to realize high-density synaptic device array for processing massive data and vertical structuring can be a viable way of achieving the goal. Synaptic transistor with vertical channel can be designed as shown in Fig. 8(a), and further, a quantum well can be equipped for low-power learning operation and effective STM [27, 28]. The potentiation is performed by band-to-band tunneling through SiGe with higher power efficiency as the simulation results in Fig. 8(b) and (c).

The valence band offset (VBO) between SiGe and Si provides a quantum well for effective hole confinement for STM as shown in Fig. 9(a) and (b). Since the heterostructure quantum well acts as the floating body for holes, the synaptic transistor can be fabricated on the bulk Si wafers cost-effectively. By this structuring, both area and power efficiencies are obtained at the same time. Fig. 10 depicts the modulation of synaptic weight (electrical conductance) by the number of pulses in the learning processes of the charge-trap synapse (CTS) [28]. Although the perfect linearity in weight modulation does not have to be fulfilled for off-chip learning, higher weight linearity is undoubtedly beneficial since the burdens in the peripheral circuits and supporting softwares can be greatly lessened. Further, the perfect linearity needs to be pursued in the on-chip learning neuromorphic system with full autonomy.

Nanowire synaptic transistor can be designed

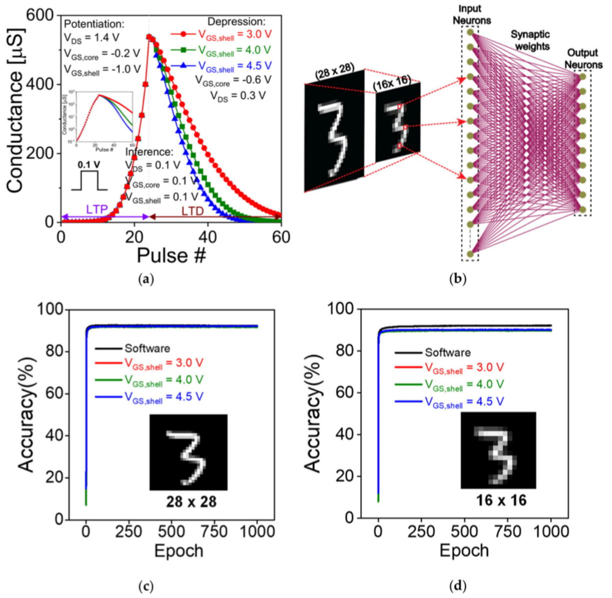


Fig. 12. Pattern recognition test of the CSDG nanowire synaptic transistor: (a) Modulation of synaptic weight in LTP and LTD characteristics of the CSDG device; (b) Schematic of the single-layer neural network made up of CSDG nanowire synaptic transistors for MNIST digit recognition. Digit recognition accuracy (%) as a functon of the number of training epochs at three different distinct depression voltages of the synaptic device for training with (c) 28 × 28; (d) 16 × 16 pixels. Insets of (c) and (d) show the MNIST images of digit “3” in the 28 × 28 and 16 × 16 resolutions, respectively [29].

considering the geometrical similarity (Fig. 11(a)) with the three-dimensional vertical NAND (VNAND) products [29, 30]. The synaptic transistor is operated by core-shell dual gates (CSDG) and the charge-trap nitride layer is located on the shell gate side as schematically shown in Fig. 11(b) and (c). Voltages of large magnitudes are applied to the shell gate for potentiation and depression operations. The core gate assists the shell gate in learning operations, being applied with voltages of smaller magnitudes. Fig. 12(a) shows the weight modulation as a function of number of learning pulses. In order to obtain higher linearity, bias conditions for potentiation, depression, and inference need to be optimized. The synaptic weights obtained from the potentiation/depression data in Fig. 12(a) was used for off-chip training of a neural network in Fig. 12(b). In comparison with the purely software-based recognition accuracy of 92.3%, there are only marginal drops in accuracy as demonstrated in Fig. 12(c) and (d), which supports the merits of the CSDG synapse.

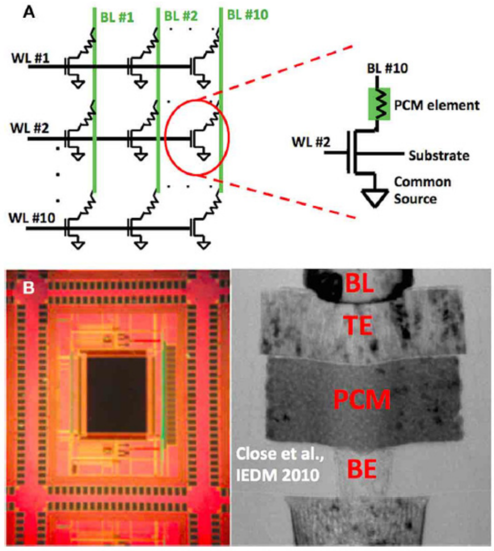


Fig. 13. Phase-change memory (PCM) synapse array: (a) Schematics of 10 × 10 array and cell; (b) Optical microscope image of PCM cell array and TEM image of a single cell [31].

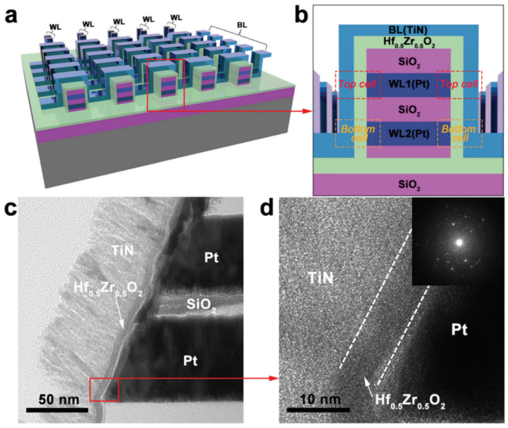


Fig. 14. 3D vertical ferroelectric HZO-based FTJ array characterization: (a) Schematic of high-density 3D vertical HZO-based FTJ synapse array; (b) zoomed-in schematic; (c) HRTEM image of the 3D TiN/FE-HZO/Pt devices; (d) Enlarged TEM image of the bottom cell corresponding to (c) (adapted from [32] with permission from Nanoscale).

2. Resistance-change Memory Synaptic Devices

As reviewed in the previous section, charge-trap flash synapses have high Si processing compatibility and can be made capable of both STM and LTM. Although the function of STM can be differed from the original one in the biological system in many aspects, one of the common functions is to manage the entire system in the energy-efficient manner. In the electronic system, sustaining the stronger connectivity with a larger weight requires a larger energy consumption. Since stronger

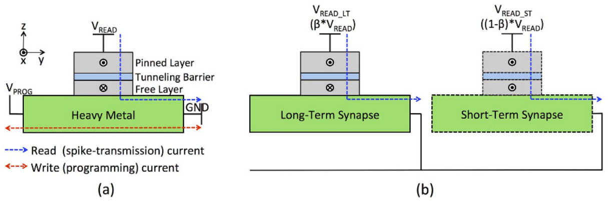


Fig. 15. Schematic of MTJ-heavy metal (HM) binary synapse: (a) Cross-sectional view; (b) A significance driven LT-ST stochastic synapse comprising two MTJ-HM devices [33].

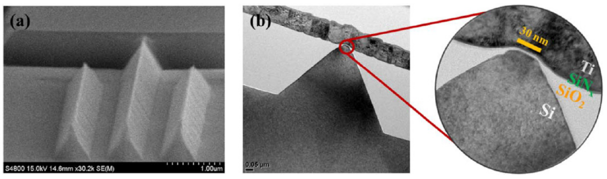


Fig. 16. (a) Construction of Si wedge. SEM image of the Si wedges for bottom electrode formation after the optimized wet etch process with 25% TMAH solution at room temperature; (b) TEM images of the cross-senctional view of Si wedge. The heavily *p*-type-doped top of the Si wedge acts as the bottom electrode of a single synaptic device cell and a bitline in the array The width of the wedge top is 30 nm (adapted from [43] with permission from Japanese Journal of Applied Physics).

connectivity implies that higher electrical conductivity, a synaptic transistor with a larger synaptic weight consumes more energy in performing inference operations at a given read voltage. From this point of view, the STM function acts as a filter discriminating less important signals - mistakenly sent signals, noises, less frequently incoming signals, etc. - that might be the sources for increasing the system power consumption by the synaptic components with unwantedly increased weights. However, STM function can be optional and can be prepared depending on system requirements and applications, and the charge-trap flash memories surveyed in the previous section can provide the plausible synaptic device solution.

As can be inferred by Fig. 11(a), synapse is the connecting part between two neurons called pre- and post-synaptic neurons. The synapse is neither an organ nor an explicit structure but an aqueous medium through which signals are propagating between the neurons. Thus, it will be closer to the reality to call it “connectivity” rather than a connecting part indeed. However, it is surely the place where two neurons meet each other so that the synapse can be treated as a two-terminal device in the electronic device sense. There can be deficiency in numbers or imperfection in functionality in realizing all the functions of a biological synapse by a device with only two terminals, and thus, assistant terminals can be added as confirmed by the charge-trap memory devices in the previous section. At the same time, a substantially large portion of researches on neuromorphic devices have been dedicated to the two-terminal synaptic devices owing to the great structural resemblance and simplicity in process integration. Resistive-switching random-access memory (RRAM), phase-change memory (PCM), ferroelectric tunnel junction (FTJ), and magnetic tunnel junction (MTJ) have been considered to be the candidates for the two-terminal synaptic devices. Fig. 13(a) and (b) through Fig. 15(a) and (b) demonstrate the synaptic devices and their arrays based on PCM, FTJ, and MTJ in the recent literature [31-33].

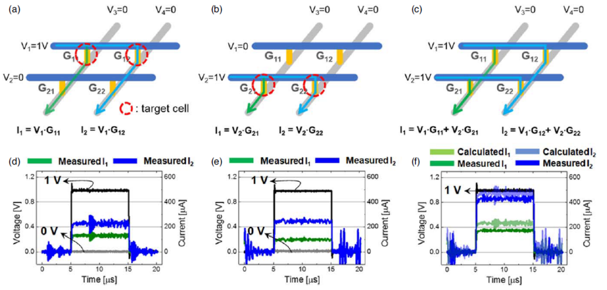


Fig. 17. Measurement for input voltage vector-conductivity matrix multiplication function of the resistive-switching synaptic device cross-point array: (a) Input voltage: (V_1, V_2) = (1 V, 0 V); (b) Input voltage: (V_1, V_2) = (0 V, 1 V); (c) Input voltage: (V_1, V_2) = (1 V, 1 V); (d) Output current for input in (a); (e) Output current for input in (b); (f) Output current for input in (c) (adapted from [43] with permission from Japanese Journal of Applied Physics).

RRAM has relatively wider variety in the base material compared with PCM, FTJ, and MTJ which usually necessitate highly delicate control over the atomic compositions. Also, RRAM has a wide span of materials compatible with Si processing, which can be a merit in the massive production point of view, including IGZO, HfO_2, TiO_x, ZTO, Ta_2O_5, SiN_x, and GeO_x can be accommodated in the contemporary Si fabrication facilities [34-42]. The resistive-switching synaptic device can be further optimized with regard to device structure for low-power operation. A novel structure of wedge can be adopted for low-voltage learning operations helped by an effective field concentration as shown in Fig. 16 [43]. The most important feature of the hardware neuromorphic system becomes apparent when the vector

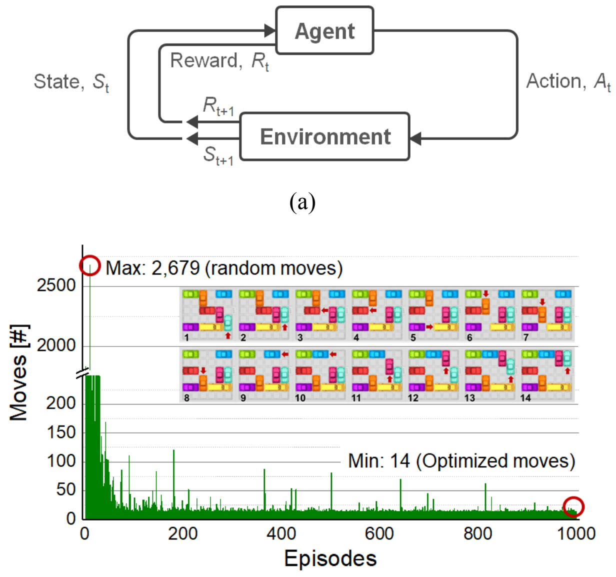
Table 1. Comparison among the reported synaptic devices

	2T DRAM [17]	SFST [24]	QW CTS (vertical) [28]	PCM [31]	HZO FTJ [32]	MTJ-HM [33]	Nanowedge RRAM [43]
Volatility	Volatile	Nonvolatile	Nonvolatile	Nonvolatile	Nonvolatile	Nonvolatile	Nonvolatile
Mechanism	Charge store	Charge trap	Charge trap	Phase change	Ferroelectric	Magnetic	Resistive-switching
Type	Charge storage type			Resistance change type			
Reported area	250 nm × 250 nm	100 nm × 100 nm	100 nm × 30 nm	90-nm node	2.5 μm^2	π/4 × 100 × 40 nm^2	30 nm × 30 nm
Processing maturity	Extremely high	Extremely high	Extremely high	Extremely high	High	High	High
Predicted cell scalability	High	High	Extremely high	Extremely high	High	Moderate	Extremely high
Multilevel operation	Possible (newly made in this work)	Possible	Possible	Possible	Possible	Possible	Possible
Switching speed	Extremely high	High	High	High	High	Extremely high	High
Inference energy	Low	Extremely low	Extremely low	Extremely low	Low	High	High

matrix multiplication (VMM) operation is clearly shown, which should be the absolute index for the accelerated MAC operations in the ultra-light and fast hardware-driven AI. Fig. 17 demonstrates the experimental results on VMM operation in the fabricated nanowedge SiN$_x$ resistive-switching synaptic device array [43].

Table 1 shows the comparison among the reported synaptic devices introduced in Chap. II and III with respect to the representative characteristics, in the order or their appearance. The weight volatility, weight modulation mechanism, and type are identified on the first three rows, which could have been understood by Fig. 1. The cell areas reported in the references are listed on the fourth row. While some of the reported synaptic devices were fabricated and their cell areas were also explicitly clarified in the references, some of them were designed by device simulation and the cell areas were estimated by a set of critical dimensions given in the references. Processing maturity means the possibility that the invented synaptic devices can be accommodated by the current fabrication technology for commercial chip production. 2T DRAM, SFST, and QW CTS are fully compatible with the Si processing. Although ferroelectric and magnetic switching materials have been actively brought into the Si processing fabrication, there is still room for expanding the variety of materials. The resistive-switching materials also have a wide span of candidates and recent materials such as SiO$_2$ and Si$_3$N$_4$ ensure the Si processing compatibility. Based on the processing maturity, cell scalability has been further predicted, beyond the reported values, in which vertical CTS, PCM, and RRAM are highly scalable. All the reported synaptic devices are capable of multilevel operations, and in particular, it has been demonstrated that a peculiarly designed DRAM can be operated with multiple weights (20 weights in the report). The highest switching speeds are found in DRAM and MTJ synaptic devices and the lowest inference energies are realized by the charge-trap synaptic devices.

The hardware-oriented neuromorphic system is under active researches and developments for the highly mobile and energy-efficient AI. However, the essense comes from the mathematical backgrounds built up by the biological analogy. MAC operation is one of the examples. In other words, there might be still room that can be filled by the software that complementarily work with the developed hardware neuromorphic system. A recent study shows that a successful encounter between the fabricated hardware neural network and software approach can increase the intelligent performances of the system. The philosophy that agent and environment interact with each other through action and reward (Fig. 18(a)) substantially reduced the minimum number of car moves that let a targeted car out of the parking lot in a shorter time (Fig. 18(b)) [44]. By the reinforcement learning in which a reward is given, the overall learning process can be shortened and it can be more effectively mimicking the way of learning in the biological system. The hardware-oriented AI would be more dependent on memory technologies which conduct the numerous



(a)

(b)

Fig. 18. Learning results: (a) Process of reinforcement learning. Agent and environment interact with each other through action and reward; (b) Number of moves required to get the red car out of the area during the reinforcement learning process (adapted from [44] with permission from IEEE Transactions on Electron Devices).

operations with superb energy efficiency in the compact hardware, being grafted with software in part for higher intelligence.

IV. Processing-in-memory (PIM)

Processing-in-memory (PIM) is one of the traditional technologies that have been developed in the very-large-scale integration (VLSI) area. The first idea came up with a terminology of logic-in-memory that features the SRAM working between the central processing unit (CPU) and slow high-density magnetic memory domain, dating back to 1970 [45]. PIM has been explictly appearing since 1990's and the majority of PIM technology is devoted by SRAM [46]. There have been similar nomenclatures that can be understood in the same meaning of PIM as shown in Fig. 19: logic-in-memory (LIM), near-memory processing (NMP), in-memory processing (IMP), memory-centric processing, etc. In short, although PIM technologies have been developed for more than half a century in the computer architecture and VLSI fields, most of the dedication has been made in reducing the physical distance between CPU and memory domain by either shortening the interconnection or introducing a new architecture topology among functional blocks. In other words, all the above technologies are realized near the memory. So, it cannot be denied that PIM has been a rather metaphorical terminology if seen from the device point of view. Coming back to the original motivation, PIM aims to get rid of the memory bottleneck or memory "wall" in the serially processing conventional computers schematically shown in Fig. 20. This should be true since the perceived speed on the end user's side is defined by the speed of communication between the processing unit and memory domain rather than the speed of processing itself. Thinking about the device scaling limit due to quantum mechanical carrier behaviors and line-and-space pitch limit capped by parasitic resistance and capacitances, further breakthrough needs to be sought with more specifically designed semiconductor memory devices for making up the PIM cells.

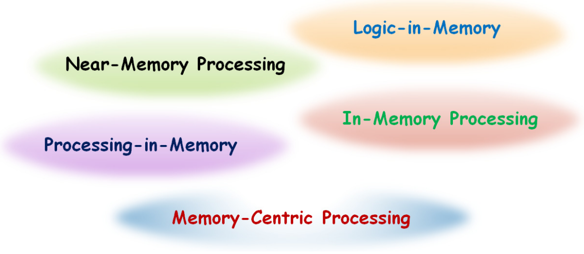


Fig. 19. Different but same names for processing-in-memory.

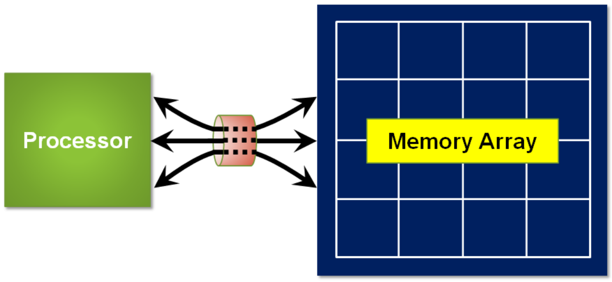


Fig. 20. Memory bottleneck in the serial-processing computers.

The understanding of difference between PIM and neuromorphic system can be helpful. Fig. 21 shows the technological map of computer architectures. Computer architectures can be categorized into Von Neumann architecture and non-Von Neumann architecture although the latter is not prevalent yet. PIM has been indicating NMP so far, indeed. Recently, a part of functions of the processing unit are allocated into individual DRAM chips, which realizes "in-memory-array" processing [47,

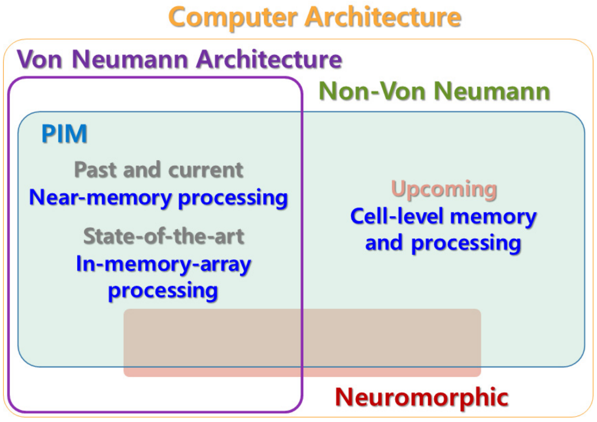


Fig. 21. Technological map of computer architectures.

48]. This is surely a new PIM technology advanced from NMP. However, the Von Neumann architecture is maintained in the in-memory-array processing. Whether the computer walks out of the Von Neumann architecture is not decidingly important if the motivation of PIM is reminded. PIM can embrace both Von Neumann and non-Von Neumann architectures only if the contributions are made in the direction of getting rid of memory wall. The literal PIM can be realized by cell-level-memory-and-processing technology, and here, the conventional architecture shall be broken. Neuromorphic computing does not completely belong to PIM but its majority is found as a subset of PIM. The reason that neuromorphic system is a subset to PIM from the task capability point of view is more succinctly glanced by the application landscape in Fig. 22. The applications can be grouped into three main categories based on the overall degree of required computational precision. A qualitative measure of the computational complexity and data accesses involved in the different applications is also shown [49]. Although neuromorhpic is mainly focused in the accerlerated MAC operations, optionally with mult-level-operational memory devices, PIM is capable of carrying out both arithmetic operations including MAC and Boolean logic operations. PIM has not existed for AI although the ingredients can make the substantial contributions. Rather, PIM is more general and universal technology in which neuromorphic can be realized as a form of PIM. Thus, indicating neuromophic system or MAC accelerator as PIM can be misleading since they take only a part in PIM. Neuromphic chip cannot replace the conventional CPU completely but PIM aims to be the new CPU technology itself. In this regard, PIM might have a new differentiating name of memory processing unit (MemPU). The final destination of logic is the memory cell itself, and at this stage, the literal PIM is realized. Breaking the Von Neumann architecture is not the goal but it can be broken at some moment while taking the forward steps to the literal PIM. It needs to be reminded that the PIM is not related with AI nor non-Von Neumann computer architecture. Not all the technologies on memory devices and integrated circuits are aiming neuromorphic system but it can be admitted that all of them are pursuing PIM for lifting up the memory wall.

V. Memory Devices for PIM Cells

SRAM and DRAM, volatile memories, showed the possibilities of implementing the cell-level memory and processing previously sketched in Fig. 21. Fig. 23 shows the in-memory computing schemes based on 8-T and 8^+-T SRAM cells in which Boolean operations of NAND, NOR, and XOR along with implication (IMP) and 2-bit read operation are realized. It is reported that 8^+-T SRAM cell in the differential mode achieves a latency of 1 ns and an average energy/bit of 29.67 fJ [50]. SRAM can tackle into PIM technology in advance due to its high operation speed but lacks of area efficiency. One of the early ideas on PIM based on volatile memory is found in the realization of in-DRAM AND and OR operations [51], which evolves into an accelerator-in-memory for bulk bitwise operations (Ambit) soon [52]. In Fig. 23, if A, B, and C represent the logical values of the three cells, then the final state of the bitline is $AB + BC + CA$ (the bitwise majority function). Since the activation is a row-level operation in DRAM, the triple-row activation (TRA) operates on an entire row of DRAM cells and multi-kilobyte-wide bitwise AND/OR of two rows is conducted [52]. Although the principles of individual memory devices are neither changed nor newly found, the full functionality for PIM can be expected when plural memory and logic devices are combined. As a result, PIM cell might be a more realistic terminology in many cases than PIM device. It would be more beneficial if the PIM technology is realized with a high capacity memory in the sense that the overall perceived speed of a system is determined by the speed of memory domain, as briefly forementioned, and the memory with the highest

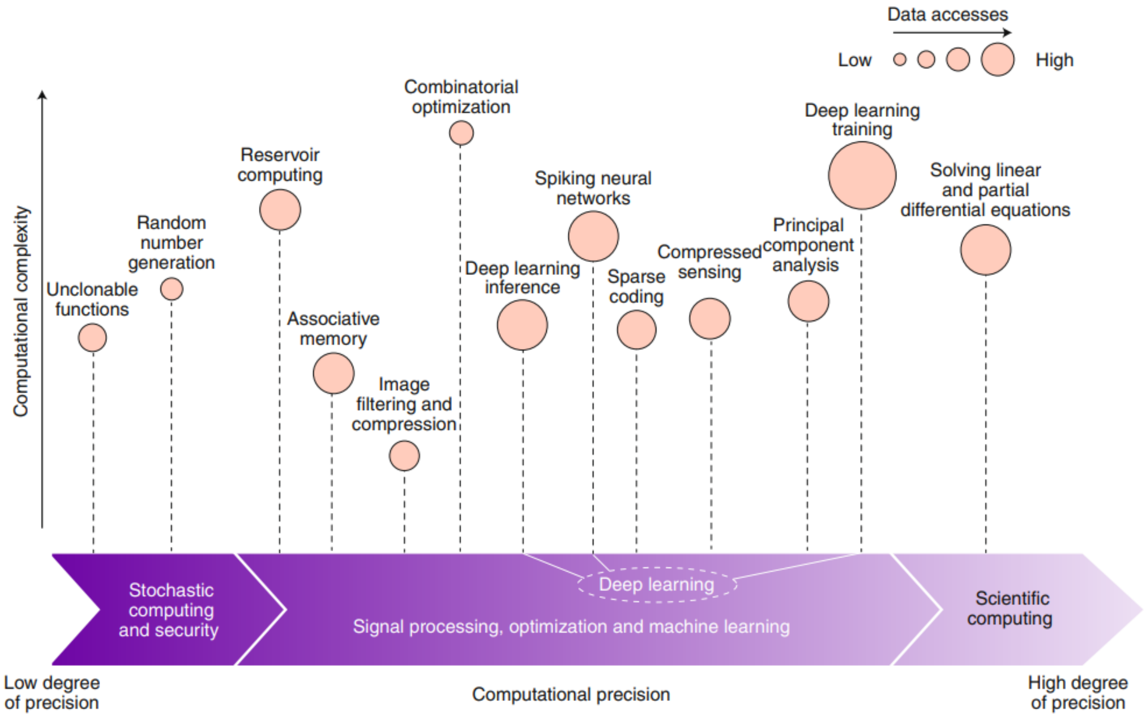


Fig. 22. Application landscape for in-memory computing (adapted from [49] with permission from Nature Nanotechnology).

density decides the eventual system speed. Thus, although the state-of-the-art PIM chip is based on DRAM at this moment [47, 48], nonvolatile memories would provide the driving force toward advanced PIM technologies just as in case of neuromorphic system. In the research level, Boolean operations are being obtained in the nonvolatile memories. Fig. 24 shows the XOR logic operation in the three-dimensional NAND flash memory array [53]. The PIM cell is implemented by a single device and the operation is conducted by the combinations of bitline and wordline voltages. PIM cell composed of two transistors and one RRAM (2T-1R) was reported [54]. Simultaneous operations of 2T-1R realizes the simultaneous logic-in-memory (SLIM) depending on the input voltages on the logic transistor gates and resistance state of the RRAM device. Fig. 25 demonstrates that NOR operation can be performed by the PIM cell as one of the feasible Boolean operations. It has been also reported that phase-change, ferroelectric, and magnetic memories can be employed in constructing a PIM cell that performs various set of Boolean operations [55-57].

Fig. 26 depicts bar diagrams to make a good distinction between NMP and the cell-level (literal) PIM. The first bar at the top shows the total sequences taken

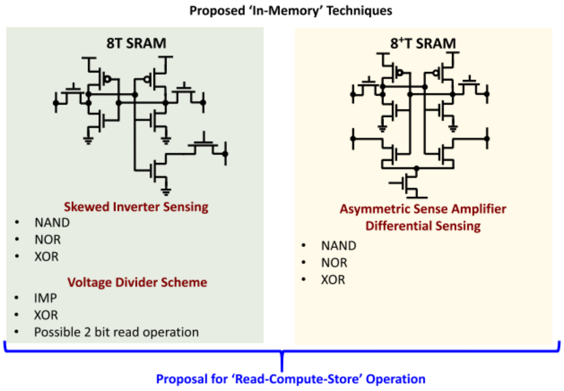

Proposed 'In-Memory' Techniques
8T SRAM
Skewed Inverter Sensing
• NAND
• NOR
• XOR
Voltage Divider Scheme
• IMP
• XOR
• Possible 2 bit read operation
8+T SRAM
Asymmetric Sense Amplifier Differential Sensing
• NAND
• NOR
• XOR
Proposal for 'Read-Compute-Store' Operation

Fig. 23. A summary of in-memory computing schemes proposed by 8-T and 8^{+}T SRAM cells (adapted from [50] with permission from IEEE Transactions on Circuits and Systems I: Regular Papers).

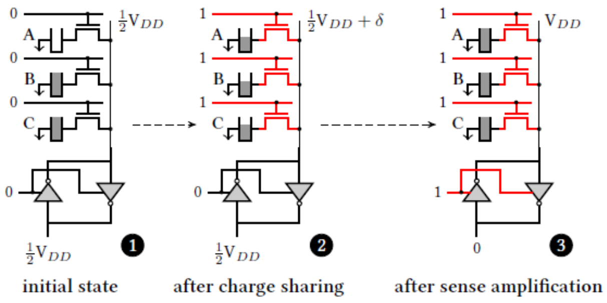


Fig. 24. Triple-row activation for in-DRAM logic operation (adapted from [52] with permission from IEEE).

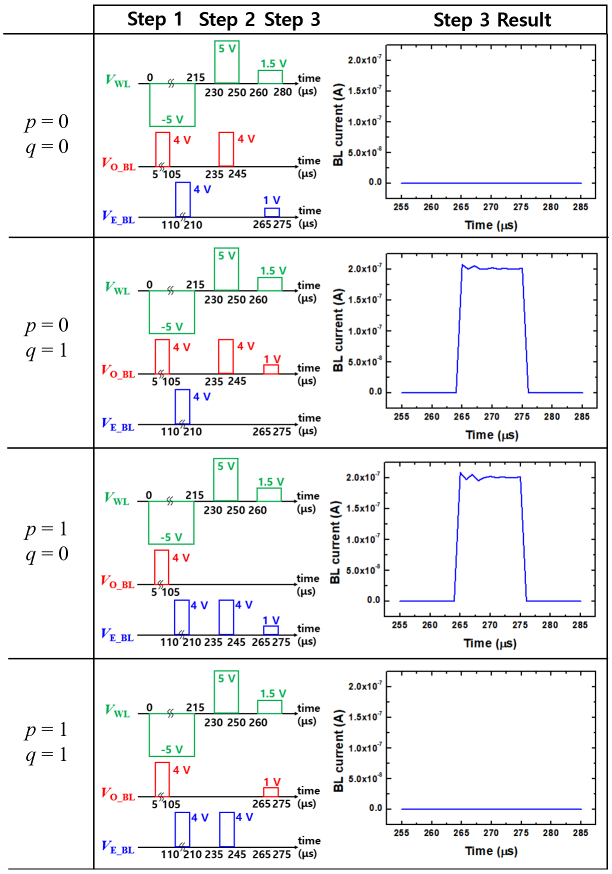


Fig. 25. Logic operations for XOR with three steps (adapted from [53] with permission from IEEE Electron Device Letters).

when CPU and memory domain communicate and the total lenth implies the time required for a unit processing/memory operation between them. Advanced computer architecture aims to reduce the length of the bar: faster logic transistor for faster CPU, interconnection with smaller RC delay, memory devices with faster read/write speeds, and effective data processing methods need to be collectively developed. These advancements result in the shorter bar at the center. The past PIM technology has dedicated to reduce the time and energy loss in the interconnection by shortening the physical distance between CPU and memory, which can be the major feature of NMP. Advanced Von Neumann architecture can be developed by minimization of the individual time segments. In-memory-array processing can be still categorized into here. On the other hand, the cell-level PIM can lift off the interconnects by specifically designed PIM cells as shown by the bar at the bottom in Fig. 26, admitting that the time save in the logic/memory devices is getting more and more

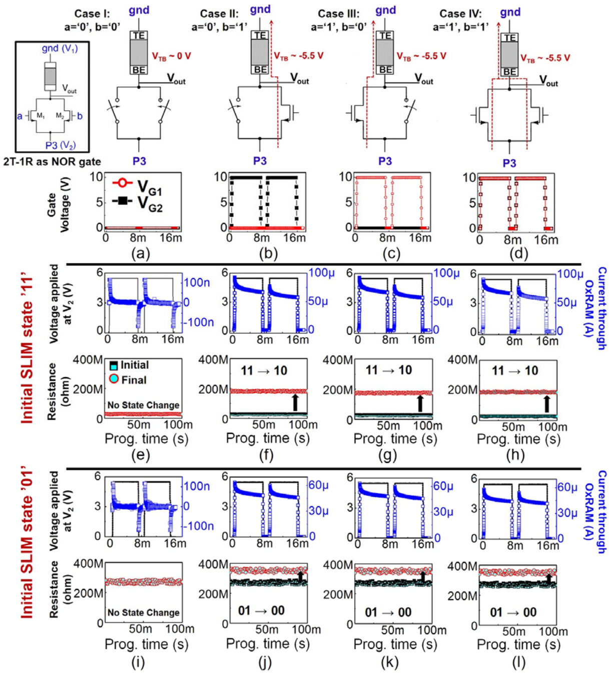


Fig. 26. Four possible input operand combinations: (a) $a = b =$ '0'; (b) $a =$ '0', $b =$ '1'; (c) $a =$ '1', $b =$ '0'; (d) $a = b =$ '1'. Experimental results for NOR logic implemented using 2T-1R SLIM bitcell with device initial state: '11' (e-h) and '01' (i-l) [54].

irreducible due to the physical and process limits. In this phase, Von Neumann architecture can be destructed.

VI. CONCLUSION

In this review, identities of volatile and nonvolatile memories have been contemplated in the view of neuromorphic and PIM technologies. Although neuromorphic system and PIM are not the same, they are not mutually exclusive at all since both of them can be implemented by memory devices. Although PIM is not targeting the AI but more widely applicable processors, both neuromorphic system and PIM resemble the human brain in which the various operations are occuring at the very place where the memory components are. Memory devices is taking the steering position for advanced computers and it should be the high time to make the series contributions toward the future processor, MemPU.

ACKNOWLEDGEMENTS

This work was supported by IITP Grant funded by the Ministry of Science and ICT (MSIT) (2021001776) and NRF Grant funded by MSIT (2021M3F3A201037927).

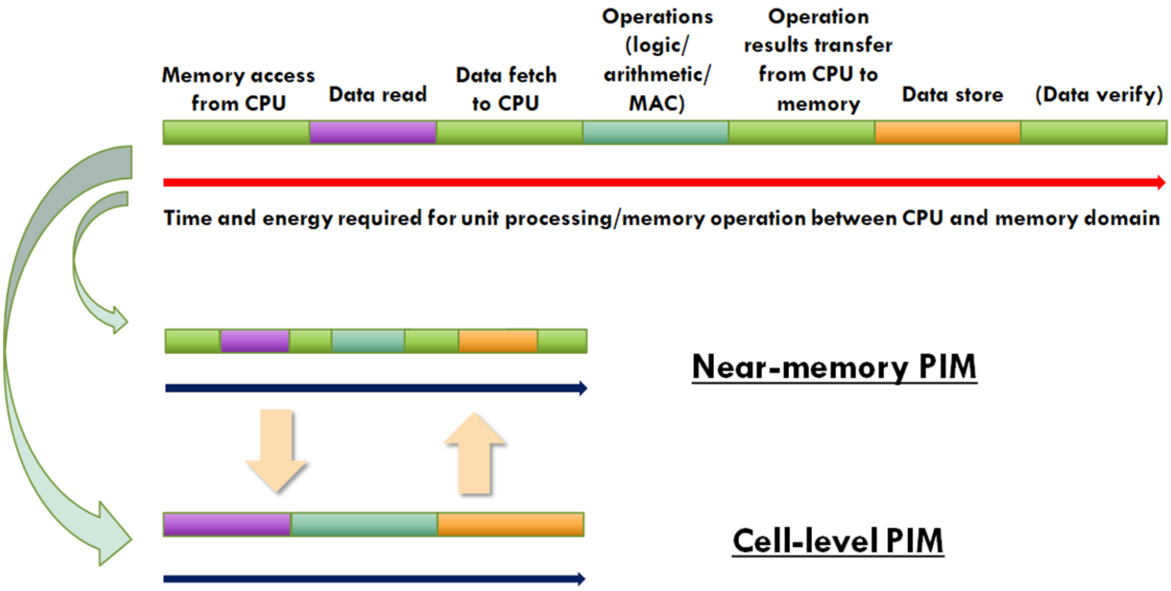

Memory access from CPU
Data read
Data fetch to CPU
Operations (logic/ arithmetic/ MAC)
Operation results transfer from CPU to memory
Data store
(Data verify)
Time and energy required for unit processing/memory operation between CPU and memory domain
Near-memory PIM
Cell-level PIM

Fig. 27. Conceptual comparison between near-memory PIM (NMP) and cell-level PIM technologies. NMP is dedicated to shorten the individual segments composing the time for whole data communiation between processor and memory domains. On the other hand, cell-level PIM can truncate one of more segments out of the entire communcation processes.

References

[1] K. Asanovic, R. Bodik, J. Demmel, T. Keaveny, K. Keutzer, J. Kubiatowicz, N. Morgan, D. Patterson, K. Sen, J. Wawrzynek, D. Wessel, and K. Yelick, “A view of the parallel computing landscape,” *Commun. ACM*, vol. 52, no. 10, pp. 56-67, Oct. 2009.

[2] V. Kindratenko and P. Trancoso, “Trends in High-Performance Computing,” *Comput. Sci. Eng.*, vol. 13, no. 3, pp. 92-95, May-Jun. 2011.

[3] K. P. Birman, “The process group approach to reliable distributed computing,” *Commun. ACM*, vol. 36, no. 12, pp. 37-54, Dec. 1993.

[4] I. Foster, Y. Zhao, I. Raicu, and S. Lu, “Cloud Computing and Grid Computing 360-Degree Compared,” *Proc. 2008 Grid Computing Environments Workshop* (*GCE*), Austin, TX, USA, 12-16, Nov. 2007. DOI: 10.1109/GCE.2008. 4738445.

[5] J. Loeffler, “AMD Zen 4 Epyc CPU could be an epic 128-core, 256-thread monster,” Techradar, Jun. 15, 2021, online available at https://www.techradar.com/news/amd-zen-4-epyc-cpu-could-be-an-epic-128-core-256-thread-monster.

[6] A. Shilov, “Arm-Based 128-Core Ampere CPUs Cost a Fraction of x86 Price,” Tom’s Hardware, Oct. 1, 2021, online available at https://www.tomshardware.com/news/ampere-altra-max-128-core-priced.

[7] Intel$^{Ⓡ}$Core^{TM}i9-10980XE Extreme Edition Processor (24.75M Cache, 3.00 GHz), online available at https://www.intel.com/content/www/us/en/products/sku/198017/intel-core-i910980xe-extreme-edition-processor-24-75m-cache-3-00-ghz/specifications.html.

[8] *Real Time Means Real Change: NVIDIA Quadro RTX 4000*, online available at https://www.nvidia.com/content/dam/en-zz/Solutions/design-visualization/quadro-product-literature/quadro-rtx-4000-datasheet.pdf.

[9] S. Cho, *Semiconductor Memory Devices for Hardware-Driven Neuromorphic Systems*, MDPI Books, Sep. 2021.

[10] C. Mead, “Neuromorphic Electronic Systems,” *Proc. IEEE*, vol. 78, no. 10, pp. 1629-1639, Oct. 1990.

[11] D. Silver, *et al.*, “Mastering the game of Go with deep neural networks and tree search,” *Nature*, vol. 529, pp. 484-489, Jan. 2016.

[12] D. Moore, “Neuromorphic Computing Gets Ready for the (Really) Big Time,” *Comm. ACM*, vol. 57, no. 6, pp. 13-15, Jun. 2014.

[13] 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.

[14] 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.

[15] A. G. Andreou, *et al*., "Real-time sensory information processing using the TrueNorth Neurosynaptic System," *2016 IEEE International Symposium on Circuits and Systems* (*ISCAS*), Montreal, QC, Canada, 22-25, May 2016. DOI: 10.1109/ISCAS.2016.7539214.

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

[17] 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," in *Proc. 2021 IEIE Summer Conf.*, pp. 515-516, Jeju, Korea, Jun. 30 - Jun. 2, 2021.

[18] S. Cho and S. Baek, "Two-Transistor Memory Cell, Synaptic Cell and Neuron Mimic Cell Using the Same and Operation Method Thereof," *Korean Patent filed*, 10-2021-0150751, Nov. 4, 2021.

[19] A. Wingfield nand D. L. Byrnes, "Decay of Information in Short-Term Memory," *Science*, vol. 176, no. 4035, pp. 690-692, May 1972.

[20] E. Camina and F. Güell, "The Neuroanatomical, Neurophysiological and Psychological Basis of Memory: Current Models and Their Origins," *Front. Pharmacol.*, vol. 8, pp. 438-1-438-16, Jun. 2017.

[21] M. M. Botvinick and D. C. Plaut, "Short-Term Memory for Serial Order: A Recurrent Neural Network Model," *Psychol. Rev.*, vol. 113, no. 2, pp. 201-233, Apr. 2006.

[22] J. Liu, H. Zhang, T. Yu, D. Ni, L. Ren, Q. Yang, B. Lu, D. Wang, R. Heinen, N. Axmacher, and G. Xue, "Stable maintenance of multiple representational formats in human visual short-term memory," *PNAS*, vol. 117, no. 51, pp. 32329-32339, Dec. 2020.

[23] K. Ichikawa and K. Kaneko, "Short-term memory by transient oscillatory dynamics in recurrent neural networks," *Phys. Rev. Res.*, vol. 3, no. 3, pp. 033193-1-033193-9, Aug. 2021.

[24] H. Kim, S. Cho, M.-C. Sun, J. Park, S. Hwang, and B.-G. Park, "Simulation Study on Silicon-Based Floating Body Synaptic Transistor with Short- and Long-Term Memory Functions and Its Spike Timing-Dependent Plasticity," *J. Semicond. Technol. Sci.*, vol. 16, no. 5, pp. 657-663, Oct. 2016.

[25] Y. J. Lee, S. Baek, and S. Cho, "Assessment of Data Retainability in Capacitorless Dynamic Random-Access Memory by Time- and Position-Dependent Hole Diffusion Function," *2021 Asia-Pacific Workshop on Fundamentals and Applications of Advanced Semiconductor Devices* (*AWAD*), B1-6, Sandai, Japan, Aug. 26-27, 2021.

[26] Y. Cho, J. Y. Lee, E. Yu, J.-H. Han, M.-H. Baek, S. Cho, and B.-G. Park, "Design and Characterization of Semi-Floating-Gate Synaptic Transistor," *Micromachines*, vol. 10, no. 1, pp. 32-41, Jan. 2019.

[27] E. Yu, S. Cho, and B.-G. Park, "A Silicon-Compatible Synaptic Transistor Capable of Multiple Synaptic Weights toward Energy-Efficient Neuromorphic Systems," *Electronics*, vol. 8, no. 10, pp. 1102-1-1102-12, Sep. 2019.

[28] E. Yu, S. Cho, K. Roy, and B.-G. Park, "A Quantum-Well Charge-Trap Synaptic Transistor with Highly Linear Weight Tunability," *IEEE J. Electron Devices Soc.*, vol. 8, pp. 834-840, Aug. 2020.

[29] Md. H. R. Ansari, U. M. 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.

[30] Md. H. R. Ansari, S. Cho, J.-H. Lee, and B.-G. Park, "Core-Shell Dual-Gate Nanowire Memory as a Synaptic Device for Neuromorphic Application," *IEEE J. Electron Devices Soc.*, vol. 9, pp. 1282-1289, Dec. 2021.

[31] S. B. Eryilmaz, D. Kuzum, R. Jeyasingh, S. B. Kim, M. Brightsky, C. Lam, and H.-S. P. Wong, "Brain-like associative learning using a nanoscale non-volatile phase change synaptic device array," *Front. Neurosci.*, vol. 8, pp. 205-1-205-11, Jul. 2014.

[32] L. Chen, T.-Y. Wang, Y.-W. Dai, M.-Y. Cha, H.

Zhu, Q.-Q. Sun, S.-J. Ding, P. Zhou, L. Chua, and D. W. Zhang, "Ultra-low power $Hf_{0.5}Zr_{0.5}O_2$ based ferroelectric tunnel junction synapses for hardware neural network applications," *Nanoscale*, vol. 10, no. 33, pp. 15826-15833, Sep. 2018.

[33] G. Srinivasan, A. Sengupta, and K. Roy, "Magnetic Tunnel Junction Based Long-Term Short-Term Stochastic Synapse for a Spiking Neural Network with On-Chip STDP Learning," *Sci. Rep.*, vol. 6, pp. 29545-1-2954513, Jul. 2016.

[34] S. Bang, M.-H. Kim, T.-H. Kim, D. K. Lee, S. Kim, S. Cho, and B.-G. Park, "Gradual switching and self-rectifying characteristics of Cu/α-IGZO/p^+-Si RRAM for synaptic device application," *Solid-State Electron.*, vol. 150, pp. 60-65, Dec. 2018.

[35] D. K. Lee, M.-H. Kim, T.-H. Kim, S. Bang, Y.-J. Choi, S. Kim, S. Cho, and B.-G. Park, "Synaptic behaviors of HfO_2 ReRAM by pulse frequency modulation," *Solid-State Electron.*, vol. 154, pp. 31-35, Apr. 2019.

[36] T.-H. Kim, M.-H. Kim, S. Bang, D. K. Lee, S. Kim, S. Cho, and B.-G. Park, "Fabrication and Characterization of TiO_x Memristor for Synaptic Device Application," *IEEE Trans. Nanotechnol.*, vol. 19, pp. 475-480, Jul. 2020.

[37] J.-H. Ryu, B. Kim, F. Hussain, M. Ismail, C. Mahata, T. Oh, M. Imran, K. K. Min, T.-H. Kim, B.-D. Yang, S. Cho, B.-G. Park, Y. Kim, and S. Kim, "Zinc Tin Oxide Synaptic Device for Neuromorphic Engineering," *IEEE Access*, vol. 8, pp. 130678-130686, Jul. 2020.

[38] D. Kim, J. T. Jang, E. Yu, J. Park, J. Min, D. M. Kim, S.-J. Choi, H.-S. Mo, S. Cho, K. Roy, and D. Kim, "Pd/IGZO/p^+-Si Synaptic Device with Self-Graded Oxygen Concentration for Highly Linear Weight Adjustability and Improved Energy Efficiency," *ACS Appl. Electron. Mater.*, vol. 2, no. 8, pp. 2390-2397, Aug. 2020.

[39] D. Kang, J. T. Jang, S. Park, Md. H. R. Ansari, J.-H. Bae, S.-J. Choi, D. M. Kim, C. Kim, S. Cho, and D. Kim, "Threshold-Variation-Tolerant Coupling-Gate α-IGZO Synaptic Transistor for More Reliably Controllable Hardware Neuromorphic System," *IEEE Access*, vol. 9, pp. 59345-59352, Apr. 2021.

[40] U. Rasheed, H. Ryu, C. Mahata, R. M. A. Khalil, M. Imran, A. M. Rana, F. Kousar, B. Kim, Y. Kim, S. Cho, F. Hussain, and S. Kim, "Resistive switching characteristics and theoretical simulation of a Pt/α-Ta_2O_5/TiN synaptic device for neuromorphic applications," *J. Alloys Compd.*, vol. 877, pp. 160204-1-160204-10, Oct. 2021.

[41] S. Kim, S. Jung, M.-H. Kim, Y.-C. Chen, Y.-F. Chang, K.-C. Ryoo, S. Cho, J.-H. Lee, and B.-G. Park, "Scaling Effect on Silicon Nitride Memristor with Highly Doped Si Substrate," *Small*, vol. 14, no. 19, pp. 1704062-1-1704062-8, May 2018.

[42] J. Y. Lee, Y. Kim, M.-H. Kim, S. Go, S. W. Ryu, J. Y. Lee, T. J. Ha, S. G. Kim, S. Cho, and B.-G. Park, "Ni/GeO_x/p^+ Si resistive-switching random-access memory with full Si processing compatibility and its characterization and modeling," *Vacuum*, vol. 161, pp. 63-70, Mar. 2019.

[43] M.-H. Kim, S. Cho, and B.-G. Park, "Nanoscale wedge resistive-switching synaptic device and experimental verification of vector-matrix multiplication for hardware neuromorphic application," *Jpn. J. Appl. Phys.*, vol. 60, no. 5, pp. 050905-1-050905-5, May 2021.

[44] M.-H. Kim, S. Hwang, S. Bang, T.-H. Kim, D. K. Lee, Md. H. R. Ansari, S. Cho, and B.-G. Park, "A More Hardware-Oriented Spiking Neural Network Based on Leading Memory Technology and Its Application With Reinforcement Learning," *IEEE Trans. Electron Devices*, vol. 68, no. 9, pp. 4411-4417, Sep. 2021.

[45] H. S. Stone, "A Logic-in-Memory Computer," *IEEE Trans. Compt.*, vol. C-19, no. 1, pp. 73-78, Jan. 1970.

[46] M. Gokhale, N. Holmes, and K. Iobst, "Processing in Memory: The Terasys Massively Parallel PIM Array," *IEEE Comput.*, vol. 28, no. 4, pp. 23-31, Apr. 1995.

[47] UPMEM PIM Soluition: DRAM Processing Unit (DPU), *UPMEM Official website*, online available at https://www.upmem.com/technology/.

[48] HBM PIM: Memory redesigned to advance AI, *Samsung official website*, online available at https://www.samsung.com/semiconductor/solutions/technology/hbm-processing-in-memory/.

[49] A. Sebastian, M. L. Gallo, R. Khaddam-Aljameh, and E. Eleftheriou, "Memory devices and applications for in-memory computing," *Nat. Nanotechnol.*, vol. 15, pp. 529-544, Jul. 2020.

[50] A. Agrawal, A. Jaiswal, C. Lee, and K. Roy, "X-SRAM: Enabling In-Memory Boolean Computations in CMOS Static Random Access Memories," *IEEE Trans. Circuits Syst. I Regul. Pap.*, vol. 65, no. 2, pp. 4219-4232, Dec. 2018.

[51] V. Seshadri, K. Hsieh, A. Boroum, D. Lee, M. A. Kozuch, O. Mutlu, P. B. Gibbons, and T. C. Mowry, "Fast Bulk Bitwise AND and OR in DRAM," *IEEE Comput. Archit. Lett.*, vol. 14, no. 2, pp. 127-131, Jul.-Dec. 2015.

[52] V. Seshadri, D. Lee, T. Mullins, H. Hassan, A. Boroumand, J. Kim, M. A. Kozuch, O. Mutlu, P. B. Gibbons, and T. C. Mowry, "Abmit: In-Memory Accelerator for Bulk Bitwise Operations Using Commodity DRAM Technology," *Proceedings of the 50th Annual IEEE/ACM International Symposium on Microarchitecture (MICRO-50)*, pp. 273-287, Cambridge, MA, USA, Oct. 14-18, 2017.

[53] J. Lee, B.-G. Park, and Y. Kim, "Implementation of Boolean Logic Functions in Charge Trap Flash for In-Memory Computing," *IEEE Electron Device Lett.*, vol. 40, no. 9, pp. 1358-1361, Sep. 2019.

[54] S. K. Kingra, V. Parmar, C.-C. Chang, B.-Hudec, T.-H. Hou, and M. Suri, "SLIM: Simultaneous Logic-in-Memory Computing Exploiting Bilayer Analog OxRAM Device," *Sci. Rep.*, vol. 10, pp. 2567-1-2567-14, Feb. 2020.

[55] Y. Li, Y. P. Zhong, Y. F. Deng, Y. X. Zhou, L. Xu, and X. S. Miao, "Nonvolatile "AND," "OR," and "NOT" Boolean logic gates based on phase-change memory," *J. Appl. Phys.*, vol. 114, no. 23, pp. 234503-1-234503-4, Dec. 2013.

[56] M. Kim, K. Lee, S. Kim, J.-H. Lee, B.-G. Park, and D. Kwon, "Double-Gated Ferroelectric-Gate Field-Effect Transistor for Processing in Memory," *IEEE Electron Device Lett.*, vol. 42, no. 11, pp. 1607-1610, Nov. 2021.

[57] M. F. Gonzalez-Zalba, C. Ciccarelli, L. P. Zarbo, A. C. Irvine, R. C. Campion, B. L. Gallagher, T. Jungwirth, A. J. Ferguson, and J. Wunderlich, "Reconfigurable Boolean Logic Using Magnetic Single-Electron Transistors," *PLoS One*, vol. 10, no. 4, pp. 0125142-1-0125142-8, Apr. 2015.

Seongjae Cho received the B.S. and the Ph.D. degrees in electrical engineering from Seoul National University, Seoul, Republic of Korea, in 2004 and 2010, respectively. He worked as an Exchange Researcher at the National Institute of Advanced Industrial Science and Technology (AIST), Tsukuba, Japan, in 2009. Also, he worked as a Postdoctoral Researcher at Seoul National University in 2010 and at Stanford University, CA, USA, from 2010 to 2013. He joined the Department of Electronic Engineering, Gachon University, Seongnam, Republic of Korea, in 2013, where he is currently working as an Associate Professor. His current research interests include emerging memory technologies, advanced nanoscale CMOS devices, group-IV photonic devices, memory cells for neuromorphic and memory-centric processor technologies. He is a Senior Member of IEEE and a Lifetime Member of IEIE.