

In-memory computing to break the memory wall

Xiaohe Huang(黄晓合), Chunsen Liu(刘春森), Yu-Gang Jiang(姜育刚), Peng Zhou(周鹏)

Citation: Chin. Phys. B . 2020, 29(7): 078504 . **doi:** 10.1088/1674-1056/ab90e7
Journal homepage: http://cpb.iphy.ac.cn; http://iopscience.iop.org/cpb

What follows is a list of articles you may be interested in

Multistability and coexisting transient chaos in a simple memcapacitive system

Fu-Ping Wang(王富平), Fa-Qiang Wang(王发强)

Chin. Phys. B . 2020, 29(5): 058502 . **doi:** 10.1088/1674-1056/ab7e98

Investigation of gate oxide traps effect on NAND flash memory by TCAD simulation

He-Kun Zhang(章合坤), Xuan Tian(田璇), Jun-Peng He(何俊鹏), Zhe Song(宋哲), Qian-Qian Yu(蔚倩倩), Liang Li(李靓), Ming Li(李明), Lian-Cheng Zhao(赵连城), Li-Ming Gao(高立明)

Chin. Phys. B . 2020, 29(3): 038501 . **doi:** 10.1088/1674-1056/ab695f

Impact of variations of threshold voltage and hold voltage of threshold switching selectors in 1S1R crossbar array

Yu-Jia Li(李雨佳), Hua-Qiang Wu(吴华强), Bin Gao(高滨), Qi-Lin Hua(化麒麟), Zhao Zhang(张昭), Wan-Rong Zhang(张万荣), He Qian(钱鹤)

Chin. Phys. B . 2018, 27(11): 118502 . **doi:** 10.1088/1674-1056/27/11/118502

Graphene resistive random memory–the promising memory device in next generation

Xue-Feng Wang(王雪峰), Hai-Ming Zhao(赵海明), Yi Yang(杨轶), Tian-Ling Ren(任天令)

Chin. Phys. B . 2017, 26(3): 038501 . **doi:** 10.1088/1674-1056/26/3/038501

A novel circuit design for complementary resistive switch-based stateful logic operations

Xiao-Ping Wang(王小平), Lin Chen(陈林), Yi Shen(沈轶), Bo-Wen Xu(徐博文)

Chin. Phys. B . 2016, 25(5): 058502 . **doi:** 10.1088/1674-1056/25/5/058502

TOPICAL REVIEW — Physics in neuromorphic devices

In-memory computing to break the memory wall*

Xiaohe Huang(黄晓合)[1,†], Chunsen Liu(刘春森)[1,2,†], Yu-Gang Jiang(姜育刚)[2], and Peng Zhou(周鹏)[1,‡]

[1] State Key Laboratory of ASIC and System, School of Microelectronics, Fudan University, Shanghai 200433, China
[2] School of Computer Science, Fudan University, Shanghai 200433, China

(Received 5 April 2020; revised manuscript received 2 May 2020; accepted manuscript online 7 May 2020)

Facing the computing demands of Internet of things (IoT) and artificial intelligence (AI), the cost induced by moving the data between the central processing unit (CPU) and memory is the key problem and a chip featured with flexible structural unit, ultra-low power consumption, and huge parallelism will be needed. In-memory computing, a non-von Neumann architecture fusing memory units and computing units, can eliminate the data transfer time and energy consumption while performing massive parallel computations. Prototype in-memory computing schemes modified from different memory technologies have shown orders of magnitude improvement in computing efficiency, making it be regarded as the ultimate computing paradigm. Here we review the state-of-the-art memory device technologies potential for in-memory computing, summarize their versatile applications in neural network, stochastic generation, and hybrid precision digital computing, with promising solutions for unprecedented computing tasks, and also discuss the challenges of stability and integration for general in-memory computing.

Keywords: in-memory computing, non-volatile memory, device technologies, crossbar array

PACS: 85.25.Hv, 51.50.+v **DOI:** 10.1088/1674-1056/ab90e7

1. Introduction

Under the wave of artificial intelligence (AI) and 5G communication, how to process a large amount of data more efficiently is the fundamental problem of information technology.[1,2] According to the statistics from international data corporation (IDC), the total amount of global data is expected to reach to 100 ZB by 2023.[3] As this vast amount of information will be processed in the calculation unit and recorded in the memory unit, future requirements for computing and memory will be greatly enhanced. However, the distinct separation of computation and storage in the modern computing system has always been a natural disadvantage since being invented.[4] The speed difference between computation and memory is increasing, and a large amount of energy and time is wasting in the process of data transports, which is the so-called 'memory wall'[5] on account of the von Neumann bottleneck. Besides, while the feature size of silicon-based integrated circuits keeps on approaching its physical limit, the computational performance improvement caused by the size reduction becomes more and more distressed, and the heat dissipation problem caused by leakage current is more and more non-negligible.[6] For instance, Google's AI recognition network spent a cluster of 16000 processor cores on training for three days while expending 100 kW of power.[7] Therefore, how to further reduce power consumption and improve performance in future integrated circuits (IC) is a concern of researchers. Now a more common method is to use graphics processing unit (GPU)[8,9] or accelerator[10,11] to improve the ability of parallel computing, alone with increasing the bandwidth of the memory,[12,13] but with only a limited improvement in speed and energy consumption, such way does not fundamentally solve the bottleneck problem.[14]

In-memory computing, namely, computing at the site where data is stored is considered as one of the ultimate solutions.[15–19] This new computing architecture does not require data movement costs, and is expected to completely break the limitations of the memory wall by high-throughput *in situ* data processing. It was already proposed in 1969 to integrate the computational and storage functions of the chip on one unit,[20] but benefited from the monetization of Moore's law and the convenience of separate design of memory and calculator, people paid no attention beyond the von Neumann structure in those days. Only until recently in-memory logic operations[21,22] and matrix-vector multiplication (MVM)[23–26] have demonstrated potentially improved power/time efficiency that researchers are trying to explore different schemes to enable general in-memory computing for future.

Various emerging memory device technologies have been brought in computing hierarchy and shown orders of magnitude improvement in computing efficiency. Such as resistive switch devices like resistive random access memory (RRAM),[27,28] phase change memory (PCM),[29,30] and mag-

*Project supported by the National Natural Science Foundation of China (Grant Nos. 61925402 and 61851402), Science and Technology Commission of Shanghai Municipality, China (Grant No. 19JC1416600), the National Key Research and Development Program of China (Grant No. 2017YFB0405600), and Shanghai Education Development Foundation and Shanghai Municipal Education Commission Shuguang Program, China (Grant No. 18SG01).

†These authors contributed equally.
‡Corresponding author. E-mail: pengzhou@fudan.edu.cn

© 2020 Chinese Physical Society and IOP Publishing Ltd

netic tunnel junctions (MTJ),[31,32] with similar but simple three-layer structure, they all rely on the physical resistance to represent storage state. It is only necessary to apply the voltage across the terminals to change the characteristics of the material, and after the voltage is removed, the state can remain unchanged inherently, which laid the foundation for the realization of in-memory computing, providing a feasible solution for greatly improving the efficiency of MVM. The new usage of charge-based devices also provides new ideas for implementing in-memory computing, such as flash,[33] SRAM,[34] and FeFET.[35,36] These three-terminal charge-based field-effect transistors (FETs) are based on mature silicon manufacturing techniques, so they are closer to commercial availability. Meanwhile, adding *in situ* memory to the logical calculation, that is, the memory and calculation are as close as possible in the physical position, can also realize in-memory computing, reduce the data transportation cost through performing *in situ* storing and processing the captured data, and produce 'highly processed' information.[37,38]

In terms of applications, diversified memory technologies can perform specific functions according to their own characteristics for various digital or analog tasks. As for analog applications, many researchers have realized the simple neural network calculation by constructing a crossbar array,[39,40] realizing letter recognition,[41] image classification,[42] sparse coding,[43] etc. Analog neural network computing also provides new computational paradigms like neuromorphic computing,[44–46] which is expected to perform the brain-like synaptic function to mimic the human brain and is one of the long-term goals of chip manufacturing. Randomness is another potential advantage of in-memory computing for stochastic applications.[47–49] As for hybrid precision applications, in-memory binary computing can be easily achieved by resistive switching with lower energy and area consumption,[50,51] while more degrees of states further allow the accumulative computing application.[52,53] Although some of these studies are still in the preliminary stage, we still see the advantages and characteristics of in-memory computing in vectormatrix multiplication and brain-like computation. We should also be aware that there are still many shortcomings in the current in-memory computing technology,[18,19] such as the problem of device stability, large-scale integration, more applicable and mature algorithm. In order to implement high-performance computing for the benefit of human beings, it is necessary to overcome these difficulties and challenges.

This review focuses on the current situation of cutting-edge research on in-memory computing technologies, explores the problems encountered in large-scale fabrication and applications, and evaluates the possible solutions. In Section 2, the memory device technologies and principles used in in-memory computing will be introduced and analyzed elaboratively. Section 3 mainly talks about how in-memory computing can be realized at different application levels in terms of neural network, stochastic generation, and hybrid precision digital computing. In Section 4, we focus on the main problems and challenges encountered in the further development and give the corresponding possible solutions. In Section 5, we summarize the full text and look forward to the trend of in-memory computing.

2. Memory device technologies for computation

The implementation principle of in-memory computing is usually determined by the underlying basic unit. In general, in-memory computing needs a memory portion to store information, so it has certain commonalities with many other memory technologies.[54,55] By far, most of in-memory computing methods are developed and modified from the mature memory devices through adding computational functions. Among these devices that can perform in-memory computing, we can divide them into three types (Fig. 1): one is the emerging two-terminal resistive switch memory technologies including RRAM, PCM, and MRAM. Although they are not yet widely available in general applications, their advantages such as simple structure, easy operation, and strong functionality bring a lot of possibilities for memory integration while their large scale commercial productions have also been demonstrated.[56–58] The devices built on a conventional three-terminal transistor like flash, SRAM, and FeFET rely on the electric field to manipulate the charge carriers, so they are treated as charge-based memory which can perform calculations with only minor changes in off-the-shelf silicon-based processes. The third is *in situ* memory based on entirely new material devices, such as carbon nanotubes and two-dimensional material transistors. Thanks to the smooth lattice interface of the new materials in nanoscale, the *in situ* storage is achieved effortlessly by adding memory cells on top of the new type of transistors.

2.1. Resistive switch memory

The emerging non-volatile memory, mainly depending on the physical state changes of specific materials to complete the representation of information, is different from the traditional electronic devices that control charges. The information they store can be reflected by resistance, phase, magnetism, etc. The external performance of these devices, namely, the current response, is different under the action of the electric field but can be attributed to the change in resistance state, so they are also collectively referred to as resistive switch devices.[16]

RRAM, also known as memristor in other literature, has been first reported as early as in the 1960s as the reversible resistive effect induced by the electric pulse.[61,62] However, due to the limitations of semiconductor technological conditions and market demand at that time, it did not receive widespread attention. In 1971, Chua proposed a theoretical model of memristor,[63] which predicted the existence of such a device

whose resistance state will change with the history of the applied voltage. In 2008, HP company prepared an RRAM device based on TiO_2 in experiments,[59] and firstly connected RRAM with memristor. Usually using a metal–dielectric–metal sandwich structure, as shown in Fig. 1(a), the insulation medium in RRAM can be invertible between high and low resistance states under electric excitation, which are non-volatile. When the reading voltage is applied, the high/low resistance states will have low/high currents flowing through the resistance functional layer, so as to represent the storage information state '0'/'1'. The underlying physical mechanism of resistance switching is usually based on defects,[28] either anion-based oxygen vacancies[64] in the insulating layer or cation-based metal ions[65] injected from the electrodes. Under the action of forward voltage, these random defects form a conductive path through migration or diffusion and when the reverse voltage is applied, the defects will be re-distributed to break the conductive path. As shown in inset of Fig. 1(a), during the set operation, anion from the insulating oxide or cation from the electrode metal will converge into a conductive filament (CF) at a certain positive threshold voltage. The formation of CFs changes the RRAM devices from the high resistance state to the low resistance state. Under a certain negative threshold voltage, the CFs will break so that the RRAM devices change from the low resistance state to the high resistance state through the reset operation.[28] RRAM can be classified into unipolar[66] and bipolar[67] types according to the polarity of the voltage at which the resistance transition occurs. Unipolarity is caused by the Joule heat[68] prevailing at high currents to blow the conductive filaments. Meanwhile, the dominant factor in bipolar switching is the electric field.[28] Roughly all of these devices that rely on the migration and diffusion of the defects to change the state of the resistance can be regarded as RRAM.[16] Up to now, RRAM has been able to achieve a size of less than 2 nm,[69] the switching speed can be less than 1 ns, and very little energy is required for the programming operations.[70] Its rich material selection and simple device physics greatly facilitate array integration and three-dimensional stacking.[15] With the above advantages, RRAM is the most studied memory technology to explore in-memory computing, which has implemented binary and multi-bit computing in digital applications as well as neural network computing[24] and stochastics computing.[71] The switching mechanism of RRAM chemically redox reactions and migration determines its potential in stochastic applications. RRAM is the first memory hardware to be made into a crossbar array[39] to realize a neural network, which has triggered the upsurge of implementing neural network with memory technology.

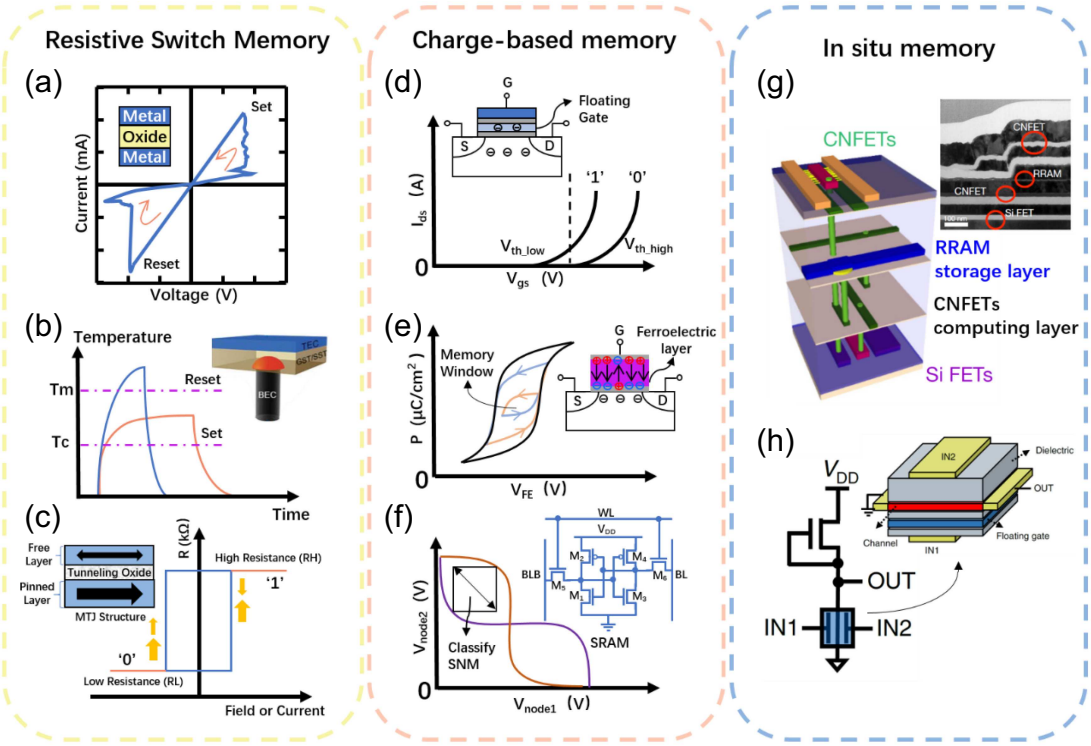


Fig. 1. Potential memory technologies for in-memory computing. (a)–(c) RRAM, PCM, and MTJ are collectively classified as resistive switch memory: (a) current–voltage characteristic and structure (inset) of an RRAM device;[59] (b) schematic diagram of programming operation and device structure (inset)[60] of PCM; (c) resistance–current characteristic and device structure (inset) of a magnetic tunnel junction. (d)–(f) Flash, FeFET, and SRAM are classified as charge-based memory: (d) transfer characteristic and device structure (inset) of a floating-gate field effect transistor; (e) polarization–voltage hysteretic characteristic and device structure (inset) of a ferroelectric field effect transistor; (f) static noise margin characteristic and unit structure (inset) of a static random access memory cell. (g) and (h) are classified as *in situ* memory, (g) the illustration of the subsystems within a three-dimensional integration chip nanosystem;[37] (h) schematic diagram of a two-dimensional dual-gate transistor with an *in situ* floating gate.[38] Panel (a) is reproduced from Ref. [59]. Panel (b) inset is reproduced from Ref. [60]. Panel (g) is reproduced from Ref. [37].

First proposed by Stanford university's Ovshinsky in the 1960s,[72] the phase change materials that make up the PCM mainly exist in chalcogenides (such as GeSbTe[73] and GeTe[74]), which can be stabilized in a polycrystalline or amorphous state. Polycrystalline is long-range ordered with low resistivity while amorphous short-range ordered with high resistivity, by which different data can be stored. The structure of PCM is shown in the inset of Fig. 1(b). The phase change material layer is sandwiched between two electrodes, and there is a heating resistance inserted between the bottom electrode and the material layer. In the set operation, a small but long-lasting current generates a high temperature through the device, increasing the temperature of the PCM layer to the crystallization temperature T_c. When the set current is kept long enough for the phase change layer to be gradually crystallization, the low resistance state will be reached after the crystallization is completed. In the reset operation, a large but short duration current passes through the heating resistor to increase the phase change material temperature above the melting point (T_m), making the material become amorphous, then it exhibits a high resistance state. The operation curve is shown in Fig. 1(b). The amorphous portion usually exhibits a "mushroom" shape (Fig. 1(b) inset), which is caused by a pillar-like heating resistor. By optimizing the structure of the heating resistor to control the Joule heat and current, the energy consumption and the miniaturization of the PCM can be further improved.[75] PCM also has the advantages of simple structure easy to scaling, robust endurance (approximately 10^9–10^{12}),[76] ultra-fast writing speed (0.7 ns),[60] and long retention time.[77,78] Also, the memory window of PCM is large enough to represent multiple states by the degree of crystallization,[79] thus realizing multiple state storage[79,80] easily. Since PCM has many similar features to RRAM, the application schemes like binary,[81,82] stochastic,[44,49] and crossbar-array[83–85] computing can also be extended to PCM. A unique feature of PCM is the multiply accumulation due to the thermally induced volume change, with which PCM has good prospects in synaptic and neuronal elements of analog artificial neural networks and multiple state accumulation.[44,86,87]

MRAM is used for information storage through a magnetic tunnel junction (MTJ).[88] The unit structure is shown in the inset of Fig. 1(c), which is mainly composed of a free layer, an insulating tunnel layer, and a pinned layer. The tunneling layer, generally MgO_2 or Al_2O_3, is extremely thin to 1 nm, to separate the free layer from the pinned layer. The magnetic domains of the pinned layer are fixed in a certain direction, and the magnetic domains of the free layer can be flipped in direction with the applied magnetic field. When the two ferromagnetic polarizations are parallel, MTJ has a low resistance. Conversely, it has high resistance. After the spin-transfer effect was proposed in 1996,[89] it has received considerable attention in MRAM again.[90,91] When the current passes through the pinned layer, the electrons whose spin direction is antiparallel to the pinned layer are reflected more significantly than the electrons whose spin direction is parallel. Consequently, the current will be polarized by spin and maintain this polarization property after passing through the tunneling layer. When arriving at the free layer, the spin angular momentum transfer torque caused by the polarization current can rotate the magnetic polarization. When the current is strong enough and lasts long enough, the magnetization of the free layer is reversed to coincide with the pinned layer. Conversely, if the current first passes through the free layer, the electrons reflected back at the surface of the pinned layer have a different polarization to the pinned layer and rotate the free layer magnetization in the antiparallel state. Therefore, we can reverse the magnetization by current to change the resistance of MTJ (Fig. 1(c)). Such a spin transfer torque mechanism can make MRAM consume less energy and smaller area[92] while have reliable writing speed as fast as 2 ns[93] and the endurance up to 10^{14},[94] with a magnetoresistance ratio (namely, the relative change of resistance) up to 200%–300%,[95] allowing STT-MRAM to be one of the most promising memory technologies. STT-MRAM can be implemented as analog synapses[96] and logic devices.[97] A stochastic MTJ inspired by the principles of neural networks reveals that it is probably the best choice to implement a true random number generator.[98,99]

2.2. Charge-based memory

Traditional flash memory is a considerably mature and highly commercial non-volatile memory that is available everywhere. Flash memory is usually divided into two categories, one is the structure of NAND, which can achieve ultra-high-density storage, but the read/write speed is slower; the other is the NOR structure, which can operate independently for each bit unit and the read/write speed is also faster, but it is relatively large. The basic unit of flash is the floating gate transistor, whose basic structure is shown in the inset of Fig. 1(d). The number of electrons in the floating gate is controlled by applying a suitable voltage pulse. In the write operation, the control gate is applied with a positive voltage, the drain terminal is applied with a high voltage pulse, and the source is grounded. The electrons in the channel get enough energy during the transport to become hot electrons scattered into the floating gate through the oxidation barrier, so that the threshold voltage is increased. In the erasing operation, the

control gate is grounded, the drain is floating, and a high voltage pulse is applied to the source, so that the electrons in the floating gate will tunnel into the source region through the oxide barrier, reducing the threshold voltage (Fig. 1(d)).[100] It is well known that the threshold voltage of the transistor can be changed by injecting or erasing charge into the floating gate, thus distinguishing the different states for storage with the advantages of low cost and high flexibility. Thanks to the maturity of physical tunneling mechanism and commercial preparation, flash can be mass-produced in large scale arrays with high accuracy, high density, high efficiency, and low cost, which is suitable for deep learning and artificial intelligence applications. The structure of the NOR flash memory array is similar to the crossbar array, making it possible for the implementation of matrix-vector multiplication.[101] Only by modifying the connecting mode of the array and adopting another read–write rules, the amount of charge in the floating gate can be precisely controlled, and flash can realize multi-bit storage. At a suitable source–drain bias, the output current of the flash will be determined by the gate voltage and the stored charge, whose accuracy can be regulated at 1%. In this way, the flash storage unit obtains the drain current through the gate voltage and the storage state, which can be regarded as a multiplication operation. A 3D NAND for computing has also been fabricated, greatly expanding the scalability of the NOR structure.[102] As long as the design of the peripheral circuit is implemented appropriately, flash can complete the function of in-memory computing efficiently.[103–105]

FeFET is a combination of traditional transistors and ferroelectric materials.[106] Due to the spontaneous polarization characteristics of ferroelectric materials, the atoms in the crystal can move over the potential barrier after applying a suitable external field (required to be larger than the coercive field E_c), leading to the inversion of the intrinsic polarization. This characteristic that the spontaneous polarization can change direction with the external electric field is called ferroelectricity. When the electric field is removed, the polarization state can be maintained so the stable passive/negative polarization states can exactly correspond to the values of '1'/'0' in binary logic, forming a binary switch.[107] The two stable states can be repeatedly reversed by the electric field, resulting in a hysteresis curve of the polarization on the electric field, as shown in Fig. 1(e), commonly referred to as the hysteresis curve, showing that the magnetic domain remains unchanged after reversal. Therefore, the ferroelectric based memory has the characteristic of non-volatile. When the ferroelectric thin film with ferroelectric polarization is introduced into the transistor gate, the threshold voltage of the transistor will be changed. Taking n-type MOS as an example (Fig. 1(e) inset), when the polarization direction is upward, negative polarization charge is induced on the lower surface of the ferroelectric thin film, which will cause the energy band on the surface of the channel to bend upward, resulting in the positive drift of the threshold voltage. When the polarization direction is downward, the lower surface of the ferroelectric thin film induces the positive polarization charge, which causes the channel surface energy band to bend downward, thus causing the threshold voltage to drift in the negative direction. When the data is written, under the action of positive/negative voltages, the steady-state polarization charge induced in the ferroelectric thin film causes the transistor to have high/low threshold voltages, corresponding to binary date '0'/'1'. When the data is read, the drain current under different threshold voltages can be obtained by applying a voltage on the drain of the transistor so that the ferroelectric polarization direction can be determined according to the magnitude of the drain current, thus corresponding to the stored binary value. Only by adding ferroelectric materials to the traditional transistor process, the storage function can be added based on calculation.[36,108] Due to the simple structure of the memory cell, one-transistor area, and non-destructive readout, FeFET has attracted extensive attention from researchers. As a memory device similar to flash, FeFET can also be used to construct crossbar neural networks, but with faster read/write speed.[109,110] Not only the vector-matrix multiplication can be realized naturally by FeFET, but also the STDP and LIF of the SNN can be implemented in FeFET.[35,111] FeFET has unique advantages of symmetric/linear programming in neuromorphic computation,[112] requiring only a transistor and a resistor to bionic the basic functions of neurons.[111]

SRAM memory cell has a variety of structures, such as 4T, 6T, 8T, and 12T, among them, the 6T structure can realize the better performance, thereby the cell structure generally adopts the 6T structure.[113] As shown in the inset of Fig. 1(f), the six transistors are symmetrically distributed, among which M1–M4 are bistable circuits including two cross-coupled invertors for interlocking and storing one digital signal. M5 and M6, namely, the access transistors, complete the function of connecting or disconnecting the memory unit from the peripheral circuit during the read and write operations. The access of the cell is enabled by the word line (WL). When the WL is at a high level, the access transistor is turned on, the stored value is transferred to the bit line (BL), and the inverse of the value is transmitted to the bit line bar (BLB), so the peripheral circuit reads information through BL and BLB. During the write operation, the peripheral circuit passes the voltage to BL and BLB as inputs, and after the WL is enabled, information is written to the memory cell. Although SRAM occupies a relatively large area, it is the fastest memory close to the computing unit, and is an indispensable part of

the processing chip, occupying a considerable area in the chip. In today's mature processer chips, with a little modification, the computing function like NOR and AND can be directly completed on SRAM,[114] and the weight value of ± 1 can be stored as a weight node to complete matrix operation.[34] In a crossbar array of 6T SRAM, the classify mode used for inference has a better static noise margin than the conventional SRAM mode. Dynamic random access memory (DRAM) is another volatile memory that is also available for in-memory computing such as Boolean logic.[115,116] However, there is no functional demonstration of MVM within DRAM arrays that can convince us DRAM is conducive to efficient parallel computing. While SRAM still consumes a considerable large cell area inside the most advanced chip, DRAM faces even greater scaling difficulties and costs more energy and circuitry for the refresh operation. Unlike SRAM, the technical processes make DRAM further away from the conventional computing and control unit. Combined with the fact that SRAM can be seamlessly integrated with today's most advanced standard processes for large-scale integration,[117] SRAM is the more attractive volatile memory technology for in-memory computing.

2.3. *In situ* memory

On the road of device miniaturization, people have found nanoscale materials with new characteristics, such as carbon nanotubes[118] and two-dimensional materials.[119] These nanomaterials can not only make transistors smaller in size and continue Moore's law, but also improve their energy efficiency and speed by orders of magnitude. In addition to the performance improvements, these new materials can also be vertically stacked using their structural advantages for *in situ* storage above the cell. In Fig. 1(g), an integrated system has four functional layers, including a silicon-based logic layer, a carbon nanotube computing layer, an RRAM storage layer, and a carbon nanotube sensing layer.[37] The top CNT sensing layer can measure the gas and record it in RRAM in real-time. The stored data can be compared by input the semi-adder in the logical layer of carbon nanotube, and the calculated results can be output to the latched circuit of the underlying silicon layer. For the first time, this chip demonstrates the integration of sensing, storage, and computation in a single chip without the need for off-chip data transfer. Different from through-silicon via (TSV) as a high bandwidth memory (HBM) bus,[12] such *in situ* data processing can achieve highest-throughput parallel processing, greatly reducing transmission energy consumption, while vertical stacking also saves a lot of areas. 3D chip integration is an effective way to break von Neumann bottlenecks and realize in-memory computing.

Two-dimensional material because of its extremely thin layer structure can also realize the new type of logic and memory (Fig. 1(h)).[38] As shown, since the top and bottom surfaces can be gate-controlled, just one 2D FET can be a logic gate. With such a logic gate, adding a layer of graphene as a floating gate below it can complete real-time *in situ* storage. Due to the structural advantages of the two-dimensional material, such process integration is easier than silicon in nanoscale. We can see that in the same chip, in the real-time operation, the calculation results are captured and saved, thus overcoming the bottleneck between storage and calculation to complete high-throughput information processing, which provides a new idea for in-memory computing. The van der Waals heterostructures intrinsic to two-dimensional materials also broaden a new perspective to neuromorphic computing. Because the atomic-level smooth surface and high surface-to-volume ratio enable electrostatic tunability and bionic spatiotemporal hyper-connectivity between different materials, such low-dimensional materials family brings potential neuromorphic functions to artificial synapses.[120–122] For example, a fully 2D inorganic/organic (MoS_2/PTCAD) hybrid heterojunction synapse is proposed with long-term plasticity and short-term plasticity behaviors,[123] serving as a novel multifunctional neuromorphic device.

3. Versatile applications

Benefited from the reduction of operations and the access cost between storage and calculation, in-memory computing can free up more possibilities to fulfill the unprecedented computing paradigms for versatile applications. The natural fusion of memory and computing makes the memory devices behave more like biological neurons, suitable for intelligent application by constructing the neural network. Also the intrinsic randomness inside the memory devices provides a strong guarantee for stochastic generation, which is the cornerstone of stochastic computing beyond the conventional serial computing. Moreover, in-memory computing still possesses the ability of hybrid precision digital computing to ensure sufficient accuracy of the calculation requirements.

3.1. Neural network application

As the core technology of information society, IC chips can be divided into digital chips, memory chips, and analog chips according to their function and market share. Nowadays most computing chips are adopting digital logic operations. Thanks to the development of silicon-based technology, CMOS logic chips can use simple binary signals to complete complex data processing, which can perform arithmetic operations expediently with digital logic and occupy a large share

in the computing market due to their high integration density and small size. Nevertheless, with the maturity of new device technologies such as RRAM and PCM, analog neural computing that has been neglected has also become feasible for large-scale applications. Because the process of the analog circuit continuous signals, the anti-interference ability and calculation accuracy are not as good as digital computing, but for specific algorithms, memory-based analog computing can achieve higher efficiency with neural network architecture.

Using Ohm's law for multiplication and Kirchhoff's current law for summation, vector-matrix multiplication, the foundation of neural network computing, can be easily mapped by the crossbar array, achieving in-memory analog computation.[18] In 2009, Xia proposed that switching characteristics of RRAM could also be used as a transfer switch to realize reconfigurable logic.[124] Meanwhile, he proposed that the memristor crossbar could be used as a brain-like synapse and the neuron of silicon-based transistor circuits could be combined to realize brain-like computing, bringing a new application direction to RRAM. In 2010, Lu first accomplished the concept of using nanoscale memristors as biomimetic synapses,[125] and for the first time implemented the STDP learning mechanism on memristors, which drew great attention to the construction of artificial neural networks for emerging non-volatile memory devices. The essence of artificial neural network is the parallelization of memory and computing. The advent of in-memory computing in the memristor crossbar makes the neural network available by analog computing, allowing brain-like computing to take a step further. The basic units of artificial networks are synapses and neurons.[1] Neurons, connecting by synapses with weights, can be divided into multiple layers, as shown in Fig. 2(a). These networks are usually trained by updating the synaptic weights, which can perform specific tasks such as classification after training. Implemented in the form of crossbar arrays, RRAM's input/output signals pass through nodes made up of rows and columns. Figure 2(b) is an example of an RRAM crossbar where the matrix weights are represented as conductance.[15] Unlike multiplicative accumulation (MAC), matrix-vector multiplication (MVM) of crossbar array can perform multiplication and summation in a single step by Ohm's law and Kirchhoff's law, which often takes a large amount of time and power in classical von-Neumann computers. Due to the structural advantages of resistive switch devices, high-density memory can be easily achieved by a crossbar array.[56] New non-volatile devices, including MTJ,[58] PCM,[57] etc., can build artificial neural networks by the same way. In addition to two-terminal devices, emerging three-terminal transistors such as FeFET, flash, and SRAM can also be used to build neural networks. For instance, in a FeFET formed by integrating a ferroelectric capacitor into the gate of a transistor, the degree of polarization of the ferroelectric capacitor is related to the transconductance of the channel.[36] FeFET can provide fast programming speed, low power consumption, and smooth symmetrical analog programming. Using the above crossbar array structure, FeFET can also implement matrix vector multiplication naturally, whose analog state can represent the weight of synapses in the fully connected neural network. Expect for the way how the charge is stored, flash can do exactly the same thing as FeFET with more readily manufacturing techniques. The 3D NAND flash is transformed into a accelerator for large neural networks ($>$ 100 Mb weight) with $\sim$ 40 TOPs/W.[105] Due to the precise charge control from the gate dielectric, both FeFET and flash can present symmetric/linear weight update, but the flash has a slow speed and higher energy consumption for its tunneling mechanism.

Compared with the high energy consumption of the traditional CMOS circuit, the resistive switch device is regarded as the electronic synapse to store weight and transmit signals at the same time, so it is more conducive to exploit the analog in-memory computing with ultra-low power consumption. The two-terminal resistive switch easily participates in data processing while accomplishing the data storage, providing high data throughput that is of great significance for various AI applications. Note that conventional silicon-based computing components are still required for the resistive switch device's potential commercialization in the near future, the resistive switch devices can be fabricated compatibly over the existing CMOS substrate due to the low thermal budget. A 54×108 memristor crossbar was integrated on the top of CMOS circuits including all the necessary interface circuitry, digital buses, and an OpenRISC processor (Fig. 2(c)).[23] With two write DACs, one read DAC, and a 13-bit ADC configured at each row or column of the array (Fig. 2(d)), the fully integrated chip is highly flexible to form different desired arrays (Fig. 2(e)), allowing autonomously mapping different neural network algorithms to accomplish specific computing tasks or applications. Unlike the reinforcement to the capability of access interface above, another common way to strengthen the neural network performance is by optimizing the device fabrication process to prepare more high-yield high-performance and uniform crossbar arrays. As shown in Fig. 2(f),[24] 8 uniform 2048-cells memristor arrays were also fabricated for the implementation of five-layer CNN to perform MNIST image recognition with a high accuracy of 96.19%, showcasing that deeper neural layers will significantly excavate the network's computational potential.

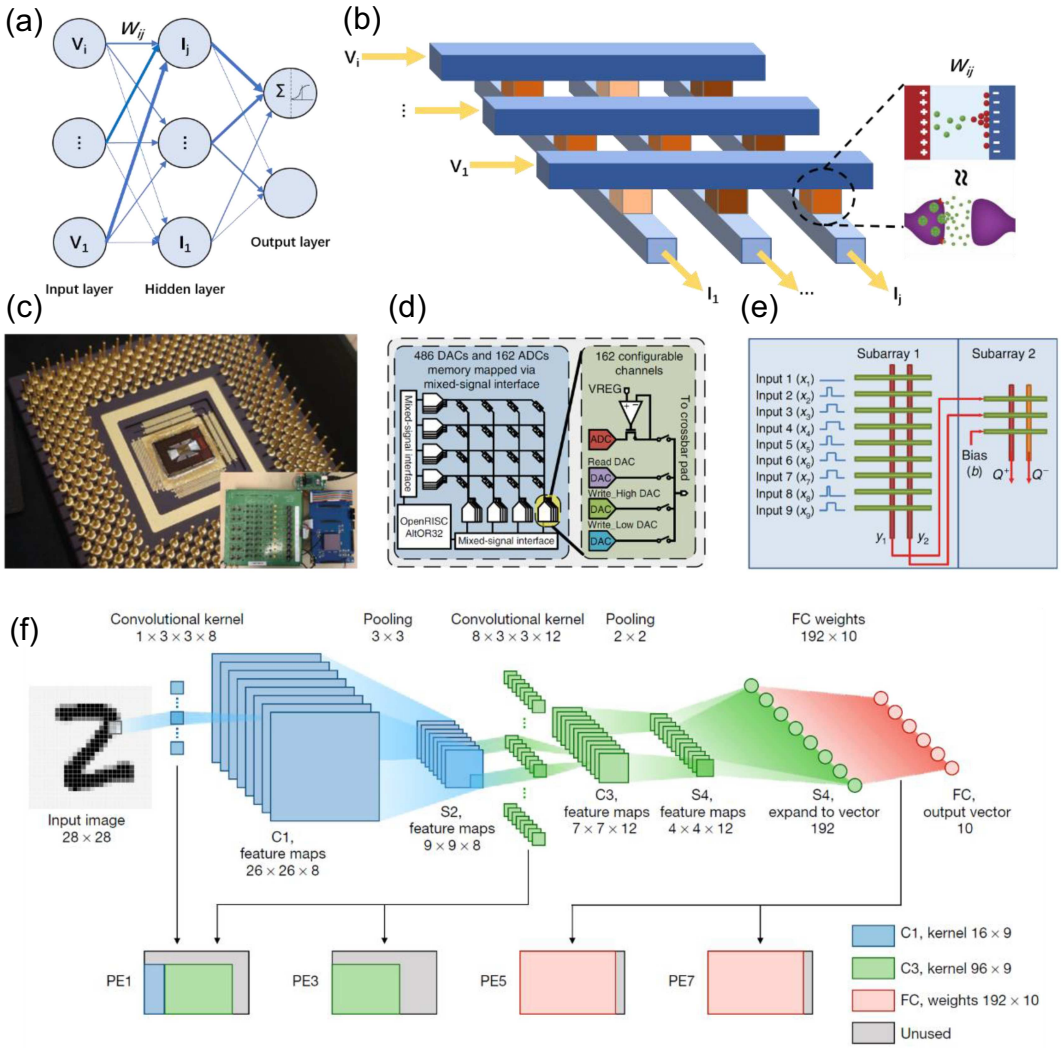


Fig. 2. Neural network demonstrations on the crossbar array. (a) Schematic diagram of a simple neural network. (b) Hardware implementation of neural network in an RRAM crossbar array. (c) A fully integrated chip with a 54 × 108 crossbar array fabricated on a CMOS substrate. (d) Schematic of the interface to the 54 × 108 crossbar array, with three read/write DACs and one ADC for each row and column. (e) The mapping of a bilayer network consisted of a 9 × 2 subarray for the PCA layer and a 3 × 2 subarray for the classification layer in the 54 × 108 crossbar array. (f) Schematic of a five-layer memristor-based CNN used for MNIST image recognition with eight 2048-cell memristor arrays. Panels (c), (d), and (e) are reproduced from Ref. [23]. Panel (f) inset is reproduced from Ref. [24].

A more biologically-inspired approach, spike neural networks (SNN),[18] can rigorously mimic the mechanism of brain information processing. CMOS circuits have been implemented in hardwares such as IBM's truenorth[126] and Intel's Loihi.[127] However, CMOS devices are ultimately unable to achieve the fusion of memory and computing, bringing a large waste of resources. It is important to find artificial synaptic devices with analog synaptic function, while simple two-terminal memristors can accomplish similar functions with obvious reduction in area, complexity, and power consumption. Artificial neurons with leaky integrate-and-fire (LIF) function have been explored by a single memristor device.[45] Furthermore, the lattice polarization dynamics of the ferroelectric layer in FeFET can imitate the learning rules of SNN, like spike-timing-dependent plasticity (STDP).[35] Novel low-dimensional materials also pave the way for SNN synaptic devices,[128] such as CNT synapses applied for unsupervised learning with an STDP scheme.[129] Figure 3 shows the timeline of neural network hardware progress through memory devices. Thanks to the efficient algorithms like the well-known back propagation, applications for DNN can be easily implemented in hardware by these new memory technologies. However, most researches on the scaling-up of SNN devices still stay in simulations. Demonstrated experiments have shown that although neuromorphic computing is at an early stage of its technological maturity without large scale implementation, it represents a promising long-term direction for AI chips and general in-memory computing.

3.2. Stochastic generation

Since the mechanism of conduction depends on the change of microscopic components, resistive switch devices such as PCM and RRAM will inevitably have certain variances that are often regarded as disadvantageous factors. In the process of IC manufacturing, we strive to reduce the vari-

ances to the smallest. However, there are great advantages in realizing the identification of specific objects by the random differences inherent in such solid-state devices. This property of randomness is almost identical to the principle of bio-identification technology, which is defined as physical unclonable function as a hardware. RRAM-based PUF, for example, not only takes advantage of the variability of integrated circuits in the manufacturing, but also uses the randomness inherent in the conductive mechanism of RRAM.[48] This inherent randomness originates from the oxygen vacancies inside the insulating dielectric between the top and bottom electrodes of the RRAM. When a conductive filament is formed and broken, the nanoscale gaps between the oxygen vacancies will inevitably change from different devices (device to device), even between different switching cycles of the same RRAM cell (cycle to cycle), and these changes can be expressed through the RRAM resistance or current. Therefore, compared with other technologies, RRAM possesses more random possibilities. The first reconfigurable PUF chip based on RRAM has been designed (Figs. 3(a) and 3(b)),[130] which has significantly improved reliability, uniformity, and chip area compared to previous work, and has unique reconfigurable capacity. SRAM-based PUF is currently the most commercially available PUF,[131] which has the advantages of easy implementation and good compatibility, but it cannot effectively resist photoelectric attack and has inherent deviation. Relatively speaking, NVM-based PUFs can effectively balance the advantages and disadvantages above, achieving low original bit error rate while possessing better resistance to physical attacks and reconfigurability. Benefitting from the lower cost, easier integration, the nanometer size, and compatibility with CMOS technology, it is a very suitable material for PUF research and has a high research value in term of hardware security that has good development potential.

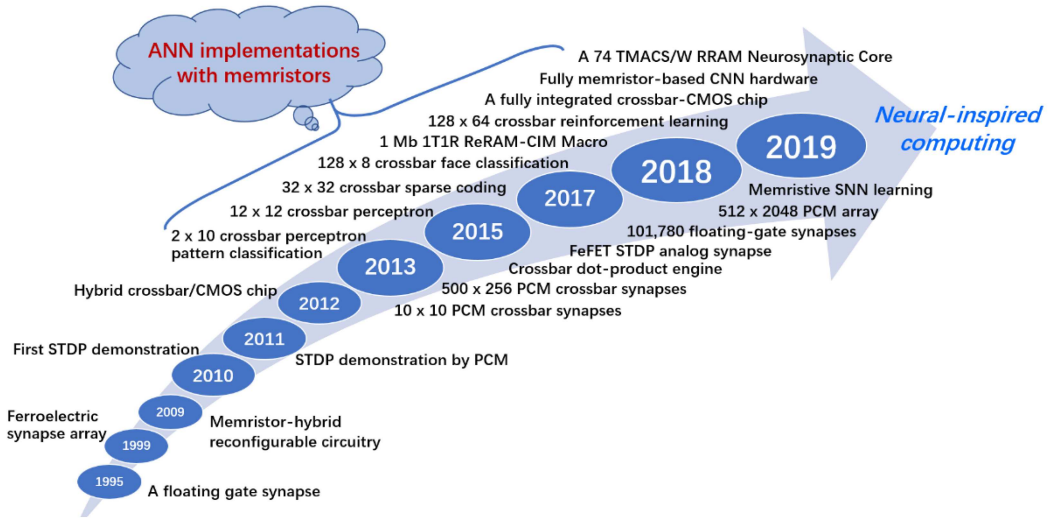
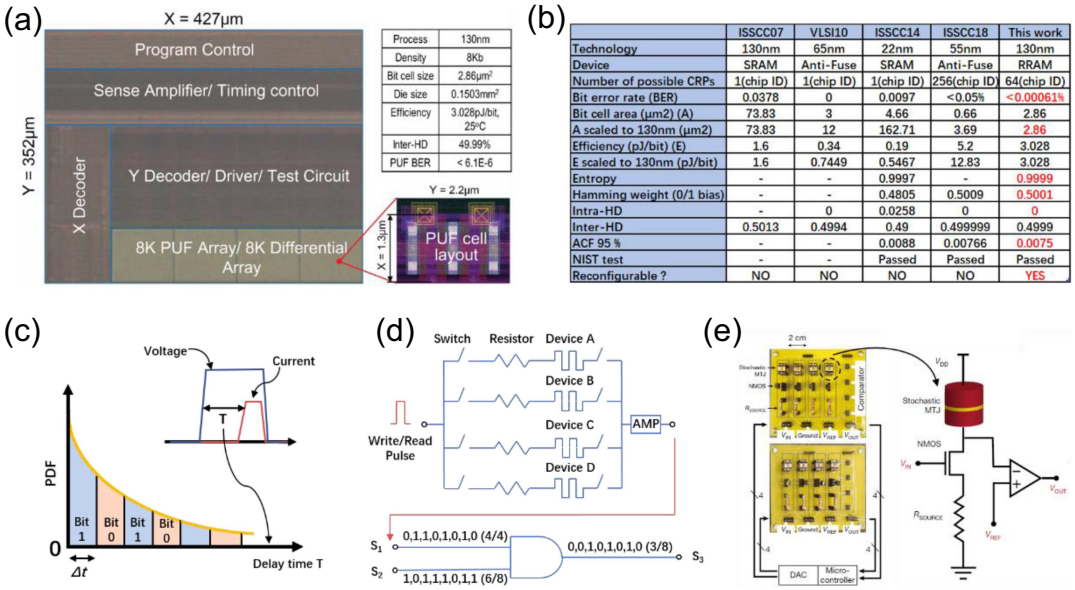


Fig. 3. Timeline of major hardware implementation on neural network towards to neural-inspired computing.

Although variability is a fatal flaw for traditional computing, but it is useful in stochastic computing, where it can be used to generate random numbers. Random numbers generations (RNG) are significantly important for stochastic computing,[132] data encryption, and neuromorphic computing.[133] Using the randomness inherent in RRAM, a true random number generator can be prepared, which is more stable and reliable than other pseudo-random number generators that need to provide seeds. Different experiments[134,135] have shown how to generate random numbers through memristors, one of which utilizes the stochastic delay time by exploiting switching variability to improve the quality of the generated random numbers (Fig. 4(c)).[136] A bias voltage slightly below the threshold value is applied across the RRAM device. After waiting for a period of time, the current through the device will suddenly increase, meaning that the RRAM is turned on, and then the device is reset back to the off state. Such a period of time for an RRAM to reset switch is stochastic, because the transition of this state is driven by thermodynamics involving the mobile transmission of oxides and ions. Theoretically, such a distribution of delay time T should satisfy the Gaussian distribution. Based on this time, we can divide by a small enough time interval Δt to guarantee a 50% probability of generating either odd or even number (corresponding to 1 or 0). A stream of random numbers can be obtained continuously by giving the same voltage and time pulse. For instance, in-memory stochastics computing only needs one AND logic gate to complete the multiplication of two fractions (Fig. 4(d)).[133,137] By using RRAM to generate random numbers, the independence and authenticity of the random numbers can be guaranteed and the accuracy of stochastic computing is improved. It also eliminates the conversion process between random numbers and binary numbers, allowing for simpler arithmetic operations that are

more efficient than traditional binary-based stochastic computing. In addition to RRAM, MRAM devices can also be used to generate random numbers with better endurance and faster switching speed. Inspired by the principle of neural network, scientists have developed a new type of hardware that enables factorization, which is one of the mathematical underlies of modern asymmetric encryption and very important for existing encryption systems. The manufacturing process and usage of such stochastics magnetic circuit hardware (Fig. 4(e))[138] based on MRAM are simpler than those for the current popular quantum computing. Now it needs only 8 bits to perform factorization on number 945. Stochastic computing has certain applications in fault-tolerant neural networks,[139] image process,[140] and decoding of check codes,[141] where NVM-based random number generation makes in-memory stochastic computing more feasible with simple circuit layout and reconfiguration.



	ISSCC07	VLSI10	ISSCC14	ISSCC18	This work
Technology	130nm	65nm	22nm	55nm	130nm
Device	SRAM	Anti-Fuse	SRAM	Anti-Fuse	RRAM
Number of possible CRPs	1(chip ID)	1(chip ID)	1(chip ID)	256(chip ID)	64(chip ID)
Bit error rate (BER)	0.0378	0	0.0097	<0.05%	<0.00061%
Bit cell area (μm2) (A)	73.83	3	4.66	0.66	2.86
A scaled to 130nm (μm2)	73.83	12	162.71	3.69	2.86
Efficiency (pJ/bit) (E)	1.6	0.34	0.19	5.2	3.028
E scaled to 130nm (pJ/bit)	1.6	0.7449	0.5467	12.83	3.028
Entropy	-	-	0.9997	-	0.9999
Hamming weight (0/1 bias)	-	-	0.4805	0.5009	0.5001
Intra-HD	-	0	0.0258	0	0
Inter-HD	0.5013	0.4994	0.49	0.499999	0.4999
ACF 95 %	-	-	0.0088	0.00766	0.0075
NIST test	-	-	Passed	Passed	Passed
Reconfigurable ?	NO	NO	NO	NO	YES

Fig. 4. Stochastic applications based on memory technologies. (a) Top view of an RRAM PUF chip with the key parameters and cell layout. (b) Comparison of performance between the RRAM PUF and other technologies for PUF. (c) Schematic of generating random numbers with the delay time from the memory devices. (d) Diagram of using one AND gate to complete fractional multiplication with generated random number streams. (e) Factorization using stochastic magnetic tunnel junctions. Left: printed circuit board of the interconnected MTJ circuit. Right: electrical schematic of a stochastic MTJ with external circuit components. Panel (a) is reproduced from Ref. [130]. Panel (c) inset is reproduced from Ref. [16]. Panel (e) inset is reproduced from Ref. [138].

3.3. Hybrid precision digital computing

Digital computing is the main function of traditional silicon-based IC chips, and its ability to solve complex computing tasks through basic digital logic has brought a lot of convenience to human life. Compared with the analog computation, published literature is scarce about its array-level experimental demonstration. In order to break through the existing bottleneck of von Neumann while ensuring the calculation accuracy, it is necessary to realize digital logic calculation for general-purpose in-memory computing. According to the number of states stored in the storage unit, the digital calculation can be divided into binary logic calculation based on switching characteristics and multi-bit calculation based on polymorphic storage. From binary to multi-bit, hybrid precision computing can be achieved in in-memory digital computing.

Since the invention of the transistor, switching transistors have a great advantage in binary computing through continual shrinking of feature size, becoming the most fundamental method of information processing technology today. In order to achieve digital calculations with low energy consumption and high area efficiency, new digital calculation concepts such as quantum dots[142] and even single atoms[143] have been proposed one after another, but these new experimental demonstrations have not yet achieved a good control of voltage and current on individual units. Instead, the new material devices like carbon nanotubes[118] and two-dimensional material[119] transistors bring out the promising performance in these respects. Profit from the flawless in the material lattice surface, both carbon nanotubes and two-dimensional material transistors can be heterogeneously integrated in nanoscale. Therefore, a memory cell is conveniently stacked on the logic transistor formed by such new materials to accomplish *in situ* memory. The three-dimensional integrated *in situ* memory chip just vertically stacks the bitwise RRAM cell on the top of carbon nanotubes field-effect transistors (CNFETs) so the half-adders constructed by the CNFETs are able to access the stored values immediately to complete conventional binary computing (Fig. 5(a)),[37] in which way the memory cost is largely reduced while maintaining the efficiency of bi-

nary computing. A novel dual-gate transistor made by two-dimensional materials is also invented for one-transistor binary logic and *in situ* memory. When both top gate and back gate are used as inputs, the charges in the floating-gate layer are attributed as the stored value to participate in the binary computing (Fig. 5(b)).[38] Such a prototype demonstration using one transistor to fulfill in-memory binary computing can achieve highest-throughput parallel processing and reduce transmission energy consumption, highlighted for the area-efficiency and multifunction.

Resistive switch devices such as RRAM and PCM show many advantages of implementing binary calculations. With the simple structure, it is easy to achieve crossbar integration and directly reconfigurable under the control of current and voltage. The high and low conductances can be used to denote the logic variables '0' and '1', respectively. More importantly, it is tantalizing for their facilitating miniaturization and non-volatile memory. Taking RRAM as an example, non-volatile binary logic can be achieved as 'stateful' logic.[50] If the resistance is also applied as the input variable so that the physical variable of input and output is uniform, such logic operation is called stateful logic. A basic stateful logic cell usually requires multiple devices to be connected in a certain way. The resistance divider between the devices will determine the final output result. There are two different stateful logic methods

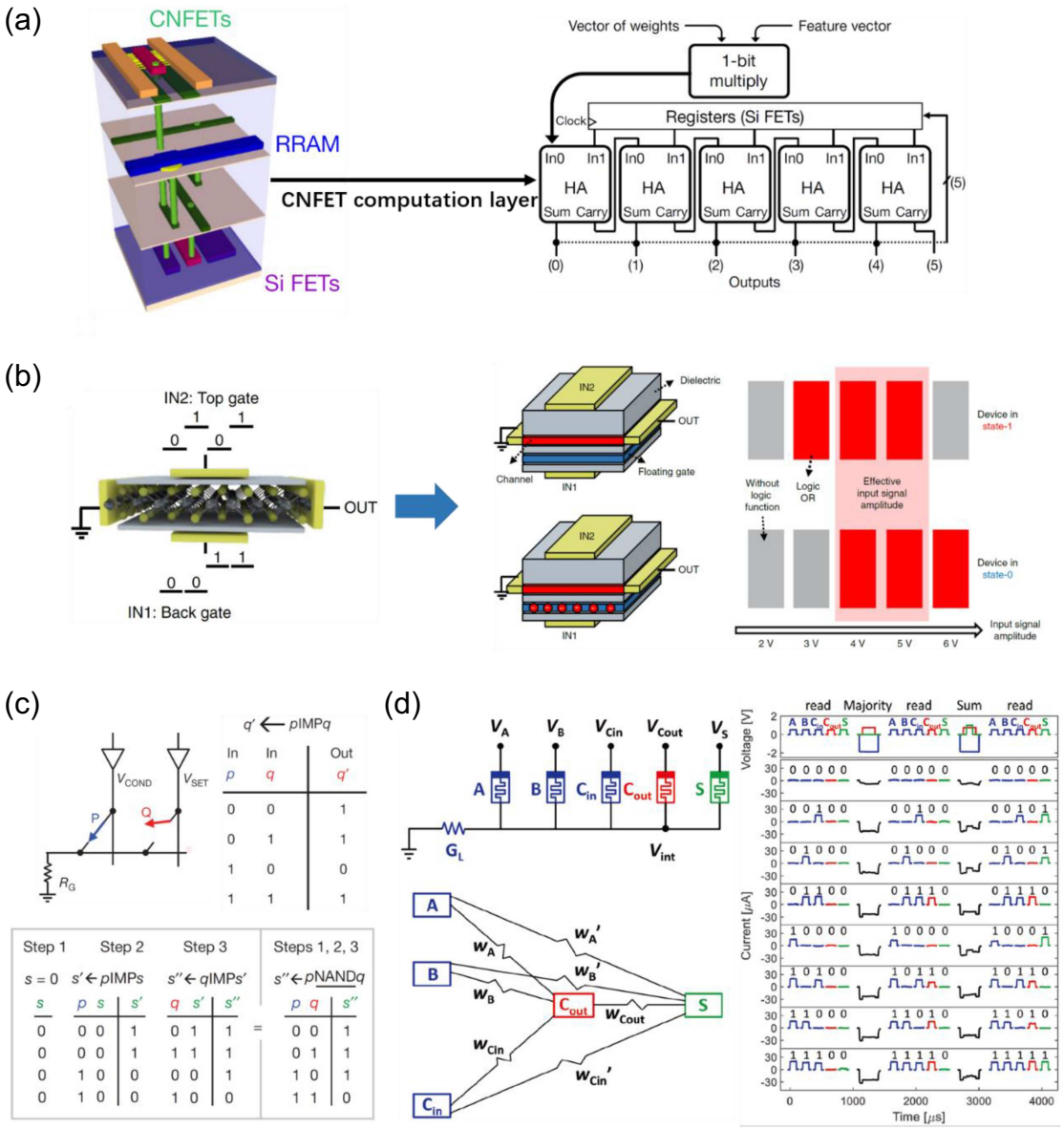


Fig. 5. Binary digital computing based on memory technologies. (a) Schematic of the classification accelerator constructed by carbon nanotubes field-effect transistors. Left: a monolithic 3D cell for the three-dimensional integration nanosystem. Right: component diagram of the CNFET computational layer. (b) *In situ* memory logic in a 2D vertical dual-gate transistor. Left: schematic of the 2D vertical dual-gate transistor structure. Right: demonstration of the logic behavior in the 2D vertical dual-gate transistor with a floating gate. (c) Stateful logic enabled by resistive switch devices. Upper: basic material implication (IMP) logic implementation. Bottom: NAND logic implementation method based on IMP and FALSE logic. (d) Full adder based on stateful logic computing of resistive switches. Left: circuit diagram and the corresponding equivalent networks of a full adder. Right: the truth table of the full adder logic. Panel (a) is reproduced from Ref. [37]. Panel (c) is reproduced from Ref. [50]. Panel (d) is reproduced from Ref. [21].

(IMP[50,144] and MAGIC[145,146]) and we take IMP for example. The IMP basic logic is composed of two identical memristors P, Q and a voltage dividing resistor R_G, as shown in Fig. 5(c). Among them, R_G is much larger than the low resistance of the memristor, but much smaller than the high resistance of the memristor. The two input variables p and q are represented by the resistance values of memristors P and Q, respectively, while the output variable q' is represented by the final resistance value of q after applying the operating voltage. Therefore, the input variables p and q are stored in devices P and Q in the form of resistance values. During the logic operation, the operating voltages V_{cond} and V_{set} are applied to the signal side of devices P and Q, respectively, and the resistor R_G is grounded. When the input variable $p = 1$, the common wire potential is about V_{cond}. At this time, the divided voltage of the device Q is not enough to change the state of the device, so the output $q' = q$. When the input variable $p = 0$, the resistance value of memristor P is high, the common wire potential is about 0, and the divided voltage of the device Q is about V_{set}. This voltage has reached the condition that the device changes from a high resistance state to a low resistance state, so the final resistance state of the device Q is low and the output result $q' = 1$. The truth table is shown in Fig. 5(c). It can be seen that this logic unit implements the material implication logic (IMP). IMP and false logic can constitute a logically complete set, based on which all 16 basic Boolean logic functions can be implemented. The corresponding logical expressions need only be expressed as a combination of IMP and FALSE, and then realized through multiple cycles. In a similar way, we can also show the implementation of the NAND logic. A complex full-adder is also composed by cascading more memristors but with more complicated operations (Fig. 5(d)).[21] Although the physical quantities of stateful logic inputs and outputs are consistent and can be easily cascaded, a logic unit requires the connection of multiple devices, which causes the problem of connection disturbance. It requires higher device uniformity and more circuit elements other than memristors. Besides, the specific logic operating voltage also increases the complexity of the peripheral circuits. When the physical variables of input and output are not uniform, such logic operation is called non-stateful logic[53] (input by voltage but output by resistance). A common way of implementing non-stateful logic is based on the bipolar/complementary memristors. Using this non-stateful logic, 14 basic Boolean logics except XOR and XNOR can be implemented in 3 steps in a single memristor, but additional steps are required to rewrite the results back into the device, given the destructive reading mode. The advantage of non-stateful logic is that a single device can be seen as a logic unit to implement a complete and efficient Boolean logic, achieving in-memory computing. However, due to different physical values of input and output, the peripheral circuit is also required to convert the signal output, which greatly reduces the advantage of in-memory computing and increases the time and space complexity of logical operation.

In addition to memristors, other resistive switch devices, such as PCM[82] and STT-MRAM,[97] also can achieve similar non-volatile logic operations, which shows the universality of non-volatile logic storage and calculation. Compared with the conventional CMOS logic, non-volatile logic uses the same circuit structure to implement reconfigurable logic by applying different voltages, not only non-volatile, but also flexible reconfiguration. However, the required steps increase with the increase in functional complexity, sacrificing a certain computing efficiency and the power consumption will increase. On the other hand, the traditional silicon-based in-memory computing devices, such as flash, FeFET, and SRAM, can complete Boolean logic and save the data in one operation based on the traditional CMOS logic operation,[108,114] though still sacrifice a certain area and power consumption, which is acceptable for the energy consumption cost by data transporting. Moreover, these novel binary computation technologies are all in the preliminary research stage with many possibilities for improvement. The *in situ* storage realized by the new material carbon nanotubes and two-dimensional materials can also integrate the memory unit on the computing cell through 3D stacking without changing the logic operation of the conventional transistors, while maintaining the advantages of traditional binary logic it can also perform high-speed parallel storage, even if the related research is still in the early proof-of-concept stage.

The optimization and advancement of memory technology help to realize the storage of more states than 0 and 1, making multiple states computations possible. For devices with conventional memory function, the binary representation is naturally easy to implement because there are only two states: 0 and 1. Part of the reason for using binary is that the states of the memory are limited. The higher degree of freedom of computation puts higher requirements on the memory device, meaning that the device unit should store more states, which is different from the binary switching characteristics. At present, device technologies that can implement multiple state storage include RRAM,[45] PCM,[52] flash,[147] and so on. Usually, a suitable electrical signal such as an accumulated pulse is applied to these memory devices, and the resistance of the device evolves accordingly. By reading out the resistance value at the appropriate time, the corresponding input electrical signal calculation is done. For example, the resistance transformation of PCM is determined by the gradual crystallization of its phase transition layer in the non-crystalline state. Under the continuous stimulation of a voltage pulse, the more crystallization, the lower resistance, so it has phase change accumulation characteristics. As shown in Fig. 6(a),[53] in order to complete the decimal addition, the number of voltage pulses of the first addend is input first, followed by the sec-

ond addend pulses. The two addends are accumulated in the PCM in the form of voltage pulses. A certain number of additional pulses need to be sent until the threshold is reached by a program/verify loop. If the threshold is set at ten pulses, the number of pulses required to reach the threshold can be used to derive the decimal calculation result, for example, like $(2+5)=(10-3)=7$ (Fig. 6(a)). Based on such accumulated characteristics, the same schemes can be used for subtraction, multiplication, and division easily while storing the results at the same time.

Other fascinating applications of in-memory computing that take advantage of the accumulative behavior are prime number decomposition[53] and temporal correlations.[148,149] Finding a factor (M) of the specific number (N) is similar to the 10-based addition. First a specified resistance threshold is set after applying the number of M pulses. Second, the device needs to be RESET every time the resistance reaches the specified threshold while the total number of N pulses are applied to the device sequentially. In the end, if M is a factor of N, the final state of the device will reach the threshold, otherwise it is not. Multiple configured PCM devices can be fed in the same number of N pulses in parallel to validate more different M factors simultaneously (Fig. 6(b)).[80] In the application of temporal correlations, unsupervised learning can be achieved between binary random processes of the devices.[148] Every PCM device is treated as a single process and reset initially. When the process is 1, the corresponding device is applied to a SET pulse whose amplitude is proportional to the sum of all 1 processes in this iteration. In this way, as the iteration increases, the conductance of the device corresponding to the process with a high correlation will increase, while the conductance of the device corresponding to the process with a low correlation will remain at a low state. Finally, binary classification can be completed through such cumulative and the high conductance reflects the strong correlation result (Fig. 6(c)). An algorithm has been demonstrated on 1000000 PCM devices (Fig. 6(d)), and such binary classification not only avoids device variability and nonlinearity, but also reduces computational complexity. Binary classification can also perform operations similar to neuromorphic rules. PCM accumulates the input pulse signal each time, but when the integral reaches a certain threshold, the conductance suddenly becomes large enough to send the pulse signal, which can imitate the LIF model and hopefully realize SNN.[44,87] However, each time the PCM device reaches the pulse threshold, an additional RESET operation is required to restore it to the off state, while RRAM does not. Compared with SNN neurons composed of silicon-based CMOS, PCM and RRAM are easy to achieve miniaturization.

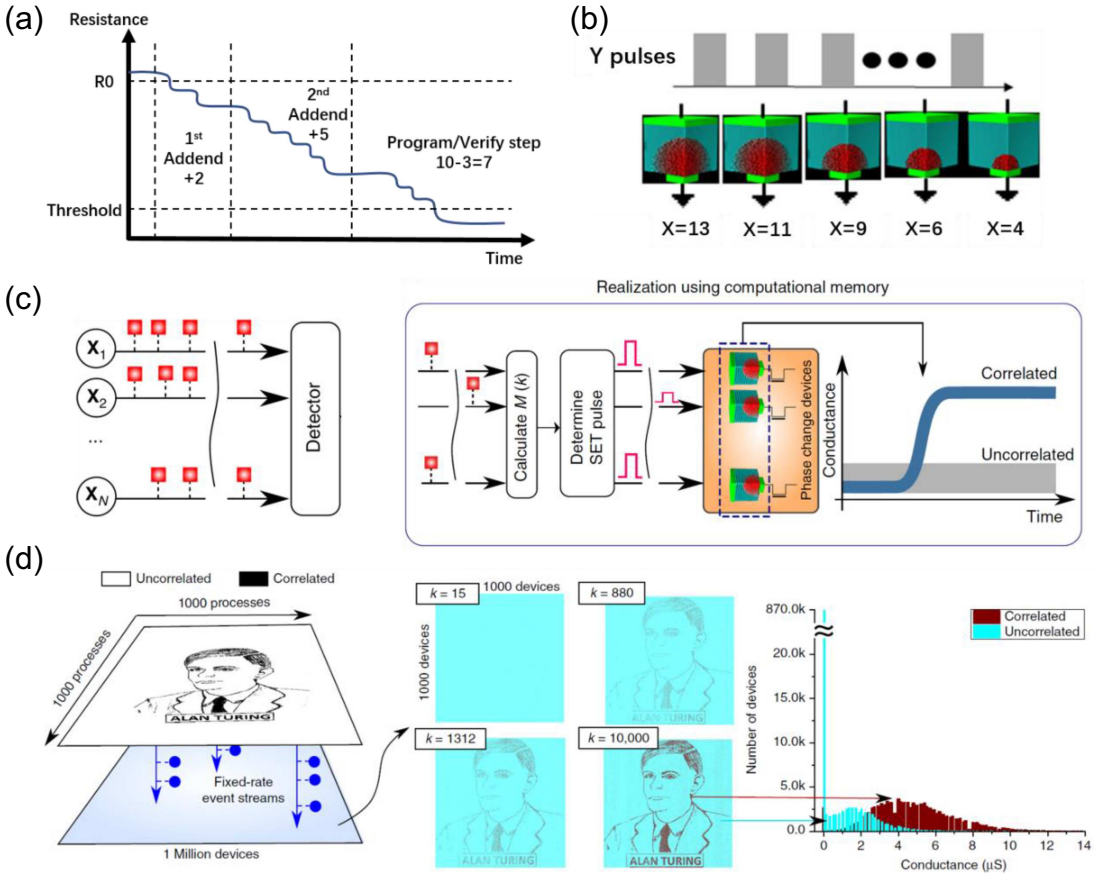


Fig. 6. Multiple-state digital computing based on memory technologies. (a) Schematic of decimal summation operation with an accumulative PCM. (b) Factorization of a number with the parallel PCM devices. (c) Flow process of temporal correlation detection achieved by accumulative PCM devices. (d) Experimental demonstration of 1000×1000 PCM devices to map the correlated processes. Panels (c) and (d) are reproduced from Ref. [148].

4. Challenges and possible solutions

The feasibility of in-memory computing cannot be separated from good device performance, large-scale integrated arrays, and applicable system-level algorithms (Fig. 7), all of which are the factors that must be considered for a universal in-memory computing in the future. Even though the promise of in-memory computing technology may seem enticing, there are still many obvious difficulties and challenges to be solved. Depending on the specific application, the problems encountered may be different. For example, the randomness of the conducting path in the memristors needs to be avoided in digital and analog computing, but no strict requirements for this in stochastic computing, instead it has important applications. In this section, we will start from the three aspects of device, array, and algorithm, analyze the problems encountered, and summarize some possible optimization solutions.

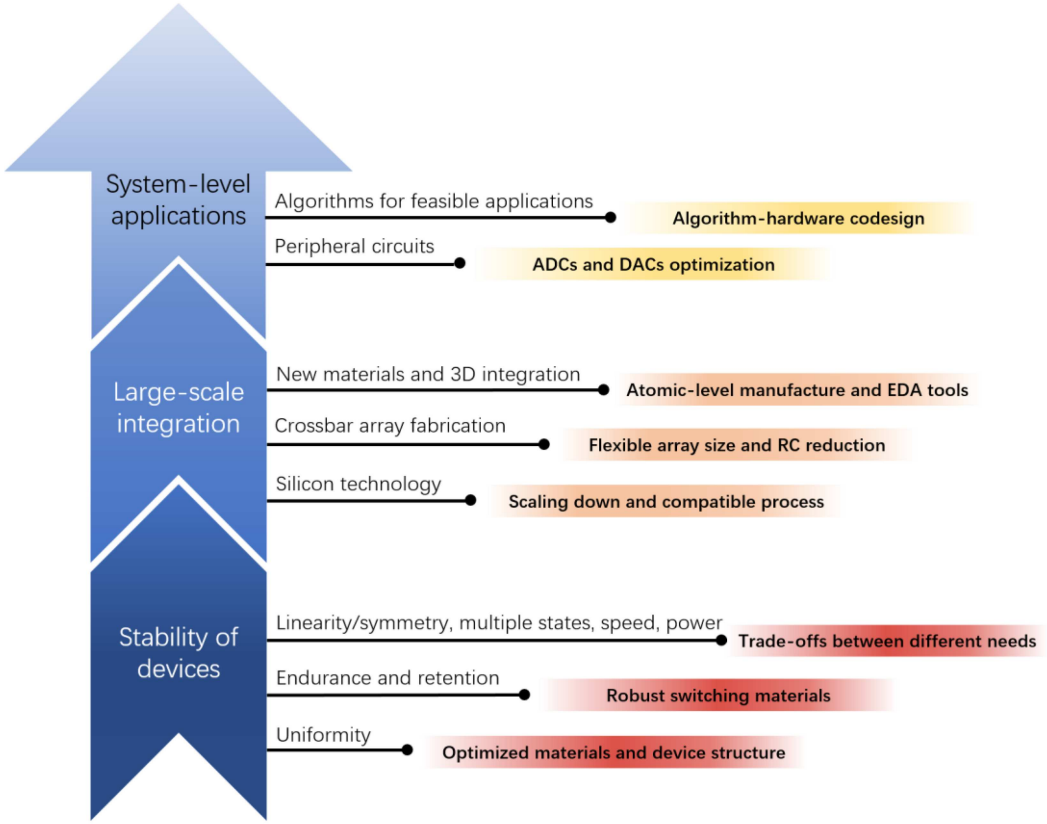


Fig. 7. The challenges and possible solutions for in-memory computing.

4.1. Stability of devices

Unlike simple binary memory, non-volatile computing places higher requirements on the performance stability of individual devices, because the requirements for the uniformity and cycle durability of the devices are higher for forcing the computation on memory. The most important issue in device research is uniformity, from device to device or cycle to cycle. Regardless of the two-terminal or three-terminal devices, the complexity and maturity of manufacture will greatly affect the uniformity between different batches even in the same batch of devices. Therefore, the technology based on the traditional silicon-based CMOS process has a great advantage on the advance of manufacture, while the new device technology is still needed to explore the more mature technologies in preparation. The difference in device mechanisms also greatly affects the issue of uniformity. For resistive switches, changes in the material properties of the resistive layers determine the transition of the device state, which is stochastics. It may be acceptable at a sufficiently large size, but due to the randomness of the ions when the device is small enough, movement or thermal activation of defects can cause fluctuations in device parameters, thereby reducing uniformity. For digital computation, the accuracy of the calculation results is very important, so the random fluctuations will reduce the calculation reliability. For analog computation, the computational disturbance caused by the randomness of microscopic particles can be used to avoid falling into the local optimal solution, so the analog computing has a certain tolerance for the variation. In particular, stochastic computing just takes good advantages of this kind of parameter fluctuation randomness to achieve truly random generation, and it is necessary to ensure that the variations are sufficiently stochastic. Of course, the parameters of each device still need to guarantee a reasonable range. The optimization for uniformity is also one of the hot research topics in the world, especially for memristors, such as using interface type memristors to replace conductive filament type memristors,[150] increasing the switching ratio to

compensate for the resistance fluctuation, introducing dislocation defects or local doping to limit the position of the shape of the conductive,[151,152] etc.

Due to the high frequency of calculation, the devices need to undergo a large number of repeated operations, which is a big test for the stability of the device. The current response during the switching process (that is, the stored procedure) inevitably produces joule heat, limiting the lifetime of the devices. The lifetime varies for different device technologies. MRAM requires only a small amount of current to operate so it can be repeated up to 10^{14} times.[94] The endurance of the memristor depends largely on the current in the working state. Uniformly based on the traditional silicon-based MOS process, the cycle durability of flash and SRAM is quite different, because flash requires large power consumption to erase and write, making its lifetime almost the shortest, while SRAM can theoretically perform countless repeated operations. In order to maximize the potential of the device, the selection of its materials and optimization of its structure are critical. For example, for memristors, selecting materials such as tantalum oxide and hafnium oxide from a resistive material system with only two stable chemical phases can effectively improve the cycling durability.[28,153] Moreover, suitable operating voltage,[154] electrode material selection,[155] and algorithm design optimization[156] are also helpful to improve endurance. A decent retention is the basic ability of these memory technologies, but in nanoscale the conductance state of the devices tends to drift with time, temperature, and voltage bias, which is prone to cause computing inaccuracy. The influence of such drift can be alleviated to a certain extent by selecting the device with a large I_{on}/I_{off} ratio.

Apart from variability and endurance, factors affecting the device include read/write speed, power consumption, number of states, symmetry, and linearity. Different device technologies have their own strengths and weaknesses in these aspects (Table 1).[157] RRAM can achieve up to hundreds of conductance states,[158] but it is not as good as PCM in terms of symmetry and linearity.[159] The volatile SRAM has a relatively large area and power consumption, but in-memory computing under the non-von Neumann structure naturally eliminates the transmission energy between memory and calculation, so the power consumption of a single SRAM cell is still acceptable, since the overall power consumption is still lower than the von Neumann architecture. A 351 TOPs/W SRAM macro was prepared in 7 nm technology node for machine-learning applications.[117] Hence, in-memory computing based on SRAM is the best choice for commercial scheme by far. Compatible with the MOS manufacturing process, flash is also more mature than other memory devices and easy to linearly program. But the major disadvantages of flash are the limited endurance, large programming power, and low speed, so flash with sufficient large storage array is suitable for the inference computing instead of the training process in the edge devices. For example, a 3D NAND flash convolutional core was proposed for deep neural networks.[33,105] Specific applications also have different requirements for devices. In digital computing like stateful logic application, fast and high endurable but low power switching characteristic is particularly demanding.[160] For neural networks applications, symmetry and linearity are especially significant in weight update to ensure the learning accuracy.[161,162] A hybrid hardware–software neural network demonstration based on PCM can only achieve the accuracy of 82% without enough symmetry and linearity while the software experiments maintain the accuracy of 97%.[161] Different memory technologies have their own strengths, so researchers should apply them with trade-off according to the appropriate needs. In order to facilitate the device design for neural network applications, a simulation platform called NeuroSim is developed,[163] which considers the non-ideal factors from device level to circuit level and benchmarks the trade-offs between different memory technologies. According to the analysis of NeuroSim, 6-bit weight and 1–2-bit weight are respectively enough to meet the needs for online training and offline inference. Endurance and variation are the main problems of these memory technologies, which need to be overcome at device level at present. Great progress can be made for in-memory computing if these two challenges are solved.

Table 1. Comparison of Pros. and Cons. between different memory technologies.

Technology	RRAM	PCM	STT-MRAM	Flash	FeFET	SRAM	*In situ* memory
Material	metal oxide	chalcogenide	magnetic	insulator	ferroelectric material	silicon	2D materials/nanotube
Switching mechanism	vacancy	Joule heating	spin injection	charge trap	electric polarization	field effect confinement	*In situ* access
Pros.	low voltage /easy structure	endurance /multi-bit	endurance /low voltage	low cost /multi-bit	speed /multi-bit	fastest	highest parallelism
Cons.	drift/variability	power	small on-off ratio	speed/endurance	instability	area/volatile	immature process

Adapted from Ref. [157].

4.2. Large-scale integration

Traditional silicon based CMOS technology has been a mainstream technology for the past 50 years due to its advantages of miniaturization for integration. Large-scale integration is the only way for in-memory computing to go beyond the laboratory and into real life applications. In-memory

computing technologies based on flash, SRAM, and FeFET are perfectly compatible with the current CMOS process, allowing the corresponding technologies to continue on the current Moore's law path. Although there is still a great debate about the slowing down or even the end of Moore's law, in-memory computing can greatly reduce the transmission power consumption for the problem of memory wall, bringing a certain increase in efficiency to the integrated chips. So exploiting the advantage of silicon-based technology is a feasible solution for large-scale in-memory computing. Using the most advanced 7 nm technology, an SRAM macro for machine learning is built with powerful computation. In addition to continued scaling to improve performance, silicon-based leading technologies can also be transferred to other memory devices' compatible processes, boosting their productivity. For example, extreme ultraviolet lithography is helpful for mass production of nanoscale cross array, while most physical vapor deposition (PVD) and atomic layer deposition (ALD) techniques can also be used to produce switching layer, dielectric layer, and metal layer with excellent conformity, accurate thickness uniformity, and component control.

Because of the simple device structure, the two-terminal resistive devices can easily be made into crossbar arrays, which have been prepared to achieve certain functions. Some of them are separate-array chips that work with external processing chips,[24] while others are manufactured by fusion with CMOS chips.[23] The current level of arrays can only deal with simple tasks. To further handle complex tasks, the size of arrays needs to be enlarged, whereas some problems will be encountered at this step. Although the crossbar array has the advantage of a simple manufacturing process, when reading the resistance value of the device, the existence of leakage current paths introduces parallel current paths, which may cause incorrect reading results and bring additional power consumption problems. Meanwhile, the crosstalk problem caused by the highly parallel writing of the array will also affect the resistance of unselected devices to a certain extent.[164] Especially for digital logic calculation, since it requires accurate reading of each independent unit, the existence of leakage current will have a great limitation on the logic operation function. All nodes need to be opened when the crossbar network is used for inference in analog computing, so the problem of leakage current does not exist, but the precise control of changing the weight of each node still needs to be avoided as much as possible during the training process. The leakage problem will become more serious as the size of the array increases, limiting the expansion of the array, and thus restricting the increase in functional complexity. The current solution to this problem includes applying a protection voltage during the operation of the array, making the memristor itself non-linear through device optimization, and striding devices in series on the memristor unit to form 1D1R, 1S1R, and 1T1R structures.[165–167] But the cascaded structure will make the unit area larger, especially the 1T1R structure. Resistive switch devices can be made to have very small feature sizes, which is also beneficial to reduce the required area to achieve the expansion of the array size, but this also introduces the problem of line resistance and capacitance. Since the interconnect metal wire is not an ideal conductor, parasitic capacitance resistance will bring RC delay and line resistance voltage division. On the one hand, it will increase the circuit delay. At the same time, the uneven voltage distribution may make the devices farther away from the power supply voltage and not work properly, resulting in a decrease in the reliability of operation.[168] In order to alleviate this problem, we can increase the width and thickness of the metal interconnect,[169] increase the access point of the power supply and ground wire,[170] exploit the novel materials like carbon nanotubes and graphene,[171] etc., all of which will of course have area consumption. Another option is to divide the array into many small arrays, connected horizontally in two dimensions or stacked in three dimensions, so that the effect of the resistance and capacitance in such small array can be eliminated. In the face of these challenges encountered in large-scale array integration, how to compromise area efficiency and computing efficiency is the main contradiction. The benefits of crossbar implemented by two-terminal devices is that it can be stacked in 3D, which undoubtedly can relax the requirement of area efficiency,[172] but it also needs to improve the preparation process.

Methods of stacking computing and memory together utilizing two-dimensional materials or carbon nanotubes are also a potential solution to the problem of memory-accessed bottleneck. By taking advantage of the fact that both the upper and lower surfaces of the two-dimensional materials can be used as input channels, and adding a floating gate as the storage layer, a transistor can be used to realize OR or AND logic gate, which has high area utilization while reducing the transport bottleneck.[38] Stanford's ILV (inter-layer via) 3D IC system is also moving beyond the university laboratory toward commercialization and large-scale application.[37] Unlike TSV 3D IC, ILV 3D IC does not stack multiple chips by packaging, but directly implements multiple chips (monolithic 3D IC) on a single wafer. Carbon nanotubes are connected to the underlying CMOS via the interlayer dielectric layer (ILD), enabling instant access to the CMOS circuit calculated results. ILV can achieve the interconnection density that 3D IC can achieve, easily reaching the feature sizes down to tens of nanometers, thus greatly improving the performance of the overall chip system. But the integration scale of these new materials is still too small for current SoCs. If carbon nanotubes and two-dimensional materials want to walk into the mainstream and promote large-scale manufacturing yield, the design of stan-

dard cell library for these new devices, as well as the EDA tools and processes,[173,174] is needed, all of which are actually a problem of design methodology and industrial ecosystems.

4.3. System-level application

To further explore the practical application of the in-memory computing, peripheral control circuits and system-level algorithms are the other two key issues that hinder the implementation of system-level chips. The peripheral circuits of most prototype concept devices are based on mature CMOS technology, which is friendly to flash, SRAM, and FeFETs compatible with silicon-based processes, but not necessarily compatible with resistive switch devices and new material devices for non-traditional manufacturing processes. Because the research of in-memory digital logic computing mainly stays at the device level, there are basically no large-scale arrays at present, and the peripheral circuit and algorithm applications at the system level are still blank, which needs more research to further realize the application. Here we mainly discuss the system-level peripheral circuits and algorithms of in-memory analog computing. Most of the research on analog computing focuses on the integration of device arrays,[25] and rarely explores the optimal design of peripheral,[23] whether it is an independent device array or an array chip integrated with silicon. Due to the need for precise programming control over the storage state (usually the conductance state) of each device, the most commonly used and important peripheral circuits are digital-to-analog converters and analog-to-digital converters.[23–25,175,176] The results of analog calculations still need to be transmitted to other digital peripheral circuits for integration processing and then feedback into the arrays, which already accounts for 60% energy consumption of the overall system, thus requiring low-power ADCs and DACs.[177] Because the matrix-vector product that originally consumed the most energy in the algorithm can now be completed in one step in the crossbar array, it seems that the multi-step processing of the peripheral circuits in the algorithm is more redundant, and the proportion of energy consumption in this part is also relatively large. Considering the various random migration processes, the device units have high mismatches in the practical manufacturing process. In order to compensate for the shortcoming of these mismatches, when completing the multi-bit calculation, additional anti-maladjustment and mismatch compensation circuits are usually required, or compensation can be performed by online learning to update neural network parameters. One solution to alleviate the requirement for peripheral circuit is to improve the robustness of the arrays. By combining high-performance crossbar arrays with a hybrid-training method, the implementation of five-layer CNN to perform MNIST image recognition with a high accuracy of 96.19% was achieved in eight uniform 2048-cells memristor arrays, providing a feasible scheme for greatly improving the efficiency of CNN (Fig. 2(f)).[24] Besides, how to use the peripheral circuit to read rows and columns efficiently also needs to be carefully designed. To better unleash the arrays' flexibility to perform MVM operations, 3 DACs and 1 ADC are interfaced with each row/column, allowing bidirectional and full transpose operation of the crossbar (Fig. 2(d)).[23] The optimization of peripheral circuits is highly dependent on the algorithm adopted by the system. For example, in the most commonly used DNN analog computing, the results obtained will select whether the external circuit is a comparator, a shift register, or a lookup table according to the step function adopted by the algorithm.[17] Therefore, the peripheral circuit needs to be combined with calculating kernels to optimize the energy efficiency and speed as a whole system. It is important to note that traditional silicon-based ADCs and DACs have also reached the saturation stage, leaving little room for energy efficiency and area optimization.

If the algorithm can reduce or optimize data process steps other than the matrix-vector product, the energy consumption and required area will be reduced even more. The closer the hardware is to the algorithm, the better the application of the algorithm will be.[19] As the most explored memory technology for in-memory computing, RRAM has been used to prepare many experimental demonstrations for artificial neural network applications from single-layer perceptron,[39,40] sparse coding[43] to reinforcement learning[25] and convolution neural network[24] (Fig. 3). Deep neural networks are now so popular because of their algorithmic applicability, where the algorithm steps can be easily mapped by a hardware that performs MVM operations excellently. A three-dimensional memristor crossbar circuit is also fabricated to further explore more parallelism and capability of memory arrays for complex neural networks.[172] The emerging SNN is closer to the human brain's thinking mode than DNN, but SNN algorithms are rarely implemented in arrays at present. Theoretically, SNN has more intelligent advantages than DNN to process voice and video information, but it is still difficult to map the algorithm to the device level.[18] Designing a hardware that cooperates with SNN rules is a big step towards neuromorphic computing, which will unleash even more unimaginably applicable computing. Unlike maturely developed DNN, SNN still lacks an efficient and streamlined hardware implementation. At least for now how to implement system-level algorithms based on SNN array chips is the key research direction.

5. Conclusion and perspectives

As an important driving force for the development of the information society in the future, the core of AI is to deal with huge amounts of data, which has led to a continuous growing search for new types of computing. In-memory comput-

ing, a non-von Neumann computing architecture, exhibits superior computational performance because it breaks through the limitations of memory wall by completely eliminating the energy and time required for data transport. Hardware demonstrations of different memory technologies have made remarkable progress (Table 2)[21,24,130,138,144,148,149,178–182] in the realization of in-memory computing as reviewed in this article. Emerging resistive switch memory plays an increasingly important role in the post-Moore era due to its simple structure and rich functional characteristics in computing. Thanks to the advanced silicon-based manufacture, charge-based memory still maintains a certain advantage in practical computing applications especially for the mature DNN algorithm. Meanwhile, *in situ* memory reveals the superiority of extreme integration in the atomic nanoscale to fulfill a multi-functional computing system. This wealth of memory technologies comes along with a wide variety of in-memory computing applications such as neural network computing, stochastic computing, and hybrid precision digital computing. The potential of neural network computing will be further released when in-memory computing can complete MVM in just one step without data movement. In-memory stochastic generation provides an almost ideal source of randomness for security and computing primitives like PUF and TRNG. Moreover, digital computing with sufficient accuracy can also be realized, which is more flexible for the characteristic of hybrid multiple-states in memory devices. In order to naturally get these applications to the practical ground, we still have to overcome the challenges from devices to algorithms levels. The stability of devices should be enhanced from optimized materials and structures while advanced and mature processes need to be explored to achieve large-scale integration. From the perspective of top-level design, the codesign of algorithm and hardware can keep the overall balance of the system and enhance the robustness, which is the key to practical application. Now we know that there are both opportunities and challenges on the road to memory integration, and how the different memory technologies are applied to in-memory computing is highly application dependent. Only to promote the advantages of respective memory technologies while restraining their drawbacks to achieve specific truly-reliable applications, can we pursue the long-term goal of general in-memory computing for the ultimate computing paradigm in the future.

Table 2. State-of-the-art implementations for applications of in-memory computing.

	Neural Network	Stochastic generation	Hybrid precision digital computing
State-of-the-art hardware implementations in terms of different applications	Stanford's 148 TOPS/W 65K RRAM neurosynaptic core[178]	TRNG: Tohoku Uni./Purdue's stochastic MJTs[138]	Binary logic: Peking Uni.'s 3M1R reconfigurable logic[144]
	Tsinghua's multiple RRAM arrays CNN system[24]	Tsinghua's analog RRAM-based TRNG[181]	POLIMI's RRAM stateful network logic[21]
	UMN's 3D NAND Flash 8-bit CNN core[179]	PUF: UCSB's 4K RRAM fixed-resistance PUF[182]	Multiple states: IBM's PCM correlation detection computing[148]
	Princeton's 658 TOPS/W SRAM CNN accelerator[180]	Tsinghua's reconfigurable RRAM PUF[130]	IBM's PCM mixed-precision computing[149]

References

[1] LeCun Y, Y Bengio and G Hinton 2015 *Nature* **521** 436
[2] Agiwal M, A Roy and N Saxena 2016 *IEEE Communications Surveys & Tutorials* **18** 1617
[3] *The Digitization of the World From Edge to Core*
[4] Neumann J V 1993 *IEEE Annals of the History of Computing* **15** 27
[5] Wulf W A and McKee S A 1995 *ACM SIGARCH Computer Architecture News* **23** 20
[6] Horowitz M 2014 *2014 IEEE Int. Solid-State Circuits Conf. Digest Tech. Papers (ISSCC)*, San Francisco, CA, 2014, pp. 10–14
[7] Le Q V 2013 *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, Vancouver, BC, 2013, pp. 8595–8598
[8] *NVIDIA Launches the World's First Graphics Processing Unit: GeForce 256* https://www.nvidia.com/object/IO_20020111_5424.html
[9] He K, Zhang X, Ren S, and Sun J 2016 *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*
[10] Chen Y H, T Krishna, J S Emer and V Sze 2017 *IEEE Journal of Solid-State Circuits* **52** 127
[11] Schmitt S, J Klän, G Bellec, *et al.* 2017 *2017 International Joint Conference on Neural Networks (IJCNN)*, Anchorage, AK, 2017, pp. 2227–2234
[12] Lee D U, Kim K W, Kim K W, Kim H, Kim J Y, Park Y J, Kim J H, Kim D S, Park H B, Shin J W, Cho J H, Kwon K H, Kim M J, Lee J, Park W, Chung B and Hong S 2014 *2014 IEEE International Solid-State Circuits Conference Digest of Technical Papers (ISSCC)*, San Francisco, CA, 2014, pp. 432–433
[13] Macri J 2015 *2015 IEEE Hot Chips 27 Symposium (HCS)*, Cupertino, CA, 2015, pp. 1–26
[14] Williams R S 2017 *Computing in Science & Engineering* **19** 7
[15] Zidan M A, Strachan J P and W D Lu 2018 *Nat. Electron.* **1** 22
[16] Ielmini D and H S P Wong 2018 *Nat. Electron.* **1** 333
[17] Yu S 2018 *Proceedings of the IEEE* **106** 260
[18] Xia Q and J J Yang 2019 *Nat. Mater* **18** 309
[19] Roy K, A Jaiswal and P Panda 2019 *Nature* **575** 607
[20] Kautz W H 1969 *IEEE Transactions on Computers* **C-18** 719
[21] Sun Z, Ambrosi E, Bricalli A and Ielmini D 2018 *Adv. Mater.* **30** e1802554
[22] Chen W H, Dou C, Li K X, Lin W Y, Li P Y, Huang J H, Wang J H, Wei W C, Xue C X, Chiu Y C, King Y C, Lin C J, Liu R S, Hsieh C C, Tang K T, Yang J J, Ho M S and Chang M F 2019 *Nat. Electron.* **2** 420
[23] Cai F, Correll J M, Lee S H, Lim Y, Bothra V, Zhang Z, Flynn M P and Lu W D 2019 *Nat. Electron.* **2** 290
[24] Yao P, Wu H, Gao B, Tang J, Zhang Q, Zhang W, Yang J J and Qian H 2020 *Nature* **577** 641
[25] Wang Z, Li C, Song W, Rao M, Belkin D, Li Y, Yan P, Jiang H, Lin P, Hu M, Strachan J P, Ge N, Barnell M, Wu Q, Barto A G, Qiu Q, Williams R S, Xia Q and Yang J J 2019 *Nat. Electron.* **2** 115
[26] Moon J, Ma W, Shin J H, Cai F, Du C, Lee S H and Lu W D 2019 *Nat. Electron.* **2** 480
[27] Waser R and M Aono 2007 *Nat. Mater.* **6** 833
[28] Yang J J, Strukov D B and Stewart D R 2013 *Nat. Nanotechnol.* **8** 13
[29] Raoux S, Welnic W and Ielmini D 2010 *Chem. Rev.* **110** 240
[30] Wong H S P, Raoux S, Kim S, Liang J, Reifenberg J P, Rajendran B, Asheghi M and Goodson K E 2010 *Proc. IEEE* **98** 2201

[31] Ikeda S, Miura K, Yamamoto H, Mizunuma K, Gan H D, Endo M, Kanai S, Hayakawa J, Matsukura F and Ohno H 2010 *Nat. Mater.* **9** 721

[32] Kent A D and Worledge H 2015 *Nat. Nanotechnol.* **10** 187

[33] Merrikh-Bayat F, Guo X, Klachko M, Prezioso M, Likharev K K and Strukov D B 2018 *IEEE Transactions on Neural Networks and Learning Systems* **29** 4782

[34] Zhang J, Wang Z and Verma N 2017 *IEEE Journal of Solid-State Circuits* **52** 915

[35] Boyn S, Grollier J, Lecerf G, Xu B, N Locatelli, Fusil S, Girod S, Carretero C, Garcia K, Xavier S, Tomas J, Bellaiche L, Bibes M, Barthelemy A, Saighi S and Garcia V 2017 *Nat. Commun.* **8** 14736

[36] Jerry M, Chen P, Zhang J, Sharma P, Ni K, Yu S and Datta S 2017 *IEEE International Electron Devices Meeting (IEDM)*, San Francisco, CA, 2017, pp. 6.2.1–6.2.4

[37] Shulaker M M, Hills G, Park R S, Howe R T, Saraswat K, Wong H P and Mitra S 2017 *Nature* **547** 74

[38] Liu C, Chen H, Hou X, Zhang H, Han J, Jiang Y G, Zeng X, Zhang D W and Zhou P 2019 *Nat. Nanotechnol.* **14** 662

[39] Alibart F, E Zamanidoost and D B Strukov 2013 *Nat. Commun.* **4** 2072

[40] Prezioso M, Merrikh-Bayat F, Hoskins B D, Adam G C, Likharev K K and Strukov D B 2015 *Nature* **521** 61

[41] Park S, Chu M, Kim J, Noh J, Jeon M, Hun Lee B, Hwang H, Lee B and Lee B G 2015 *Sci. Rep.* **5** 10123

[42] Yao P, Wu H, Gao B, Eryilmaz S B, Huang X, Zhang W, Zhang Q, Deng N, Shi L, Wong H P and Qian H 2017 *Nat. Commun.* **8** 15199

[43] Sheridan P M, Cai F, Du C, Ma W, Zhang Z and Lu W D 2017 *Nat. Nanotechnol.* **12** 784

[44] Tuma T, Pantazi A, Le Gallo M, Sebastian A and Eleftheriou E 2016 *Nat. Nanotechnol.* **11** 693

[45] Wang Z, Joshi S, Savel'ev S, *et al.* 2018 *Nat. Electron.* **1** 137

[46] Chen C, Yang M, Liu S, Liu T, Zhu K, Zhao Y, Wang H, Huang Q and Huang R 2019 *Symposium on VLSI Technology*, Kyoto, Japan, 2019, pp. T136–T137

[47] Won Ho C, Yang L, Jongyeon K, Deshpande A, Gyuseong K, Jian W and Kim C H 2014 *IEEE International Electron Devices Meeting*, San Francisco, CA, 2014, pp. 12.5.1–12.5.4

[48] Chen A 2015 *IEEE Electron Device Lett.* **36** 138

[49] Piccinini E, Brunetti R and Rudan M 2017 *IEEE Transactions on Electron Devices* **64** 2185

[50] Borghetti J, Snider G S, Kuekes P J, Yang J J, Stewart D R and Williams R S 2010 *Nature* **464** 873

[51] Linn E, Rosezin R, Tappertzhofen S, Bottger U and Waser R 2012 *Nanotechnology* **23** 305205

[52] Wright C D, Liu Y, Kohary K I, Aziz M M and Hicken R J 2011 *Adv. Mater.* **23** 3408

[53] Wright C D, Hosseini P and Diosdado J A V 2013 *Adv. Func. Mater.* **23** 2248

[54] Wong H S and Salahuddin S 2015 *Nat. Nanotechnol.* **10** 191

[55] Zhang Z, Wang Z, Shi T, Bi C, Rao F, Cai Y, Liu Q, Wu H and Zhou P 2020 *InfoMat* **2** 261

[56] Fackenthal R, Kitagawa M, Otsuka W, Prall K, Mills D, Tsutsui K, Javanifard J, Tedrow K, Tsushima T, Shibahara Y and Hush G 2014 *IEEE International Solid-State Circuits Conference Digest of Technical Papers (ISSCC)*, San Francisco, CA, 2014, pp. 338–339

[57] Kim T, Choi H, Kim M, Yi J, Kim D, Cho S, Lee H, Hwang C, Hwang E, Song J, Chae S, Chun S and Kim J 2018 *IEEE International Electron Devices Meeting (IEDM)*, San Francisco, CA, 2018, pp. 37.1.1–37.1.4

[58] Lee K, Bak J H, Kim Y J, *et al.* 2019 *IEEE International Electron Devices Meeting (IEDM)*, San Francisco, CA, USA, 2019, pp. 2.2.1–2.2.4

[59] Strukov D B, Snider G S, Stewart D R and Williams R S 2008 *Nature* **453** 80

[60] Rao F, Ding K, ZhouY, Zheng Y, Xia M, Lv S, Song Z, Feng S, Ronneberger I, Mazzarello R, Zhang W and Ma E 2017 *Science* **358** 1423

[61] Hickmott T W 1962 *J. Appl. Phys.* **33** 2669

[62] Cook E L 1970 *J. Appl. Phys.* **41** 551

[63] Chua L 1971 *IEEE Transactions on Circuit Theory* **18** 507

[64] Beck A, Bednorz J G, Gerber C, Rossel C and Widmer D 2000 *Appl. Phys. Lett.* **77** 139

[65] Liu Q, Sun J, Lv H, Long S, Yin K, Wan N, Li Y, Sun L and Liu M 2012 *Adv. Mater.* **24** 1844

[66] Guan W, Liu M, Long S, Liu Q and Wang W 2008 *Appl. Phys. Lett.* **93** 223506

[67] Ielmini D 2011 *IEEE Transactions on Electron Devices* **58** 4309

[68] Strukov D B, Alibart F and Stanley Williams R 2012 *Appl. Phys. A* **107** 509

[69] Pi S, Li C, Jiang H, Xia W, Xin H, Yang J J and Xia Q 2019 *Nat. Nanotechnol.* **14** 35

[70] Choi B J, Torrezan A C, Strachan J P, Kotula P G, Lohn A J, Marinella M J, Li Z, Williams R S and Yang J J 2016 *Adv. Func. Mater.* **26** 5290

[71] Xiao Y, Hsieh E R, Chung S S, Chen T R, Huang S A, Chen T J and Cheng O 2019 *IEEE International Electron Devices Meeting (IEDM)*, San Francisco, CA, USA, 2019, pp. 21.5.1–21.5.4

[72] Ovshinsky S R 1968 *Phys. Rev. Lett.* **21** 1450

[73] Yamada N, Ohno E, Nishiuchi K, Akahira N and Takao M 1991 *J. Appl. Phys.* **69** 2849

[74] Nakayama K, Kojima K, Hayakawa F, Imai Y, Kitagawa A and Suzuki M 2000 *Jpn. J. Appl. Phys.* **39** 6157

[75] Boniardi M, Redaelli A, Cupeta C, Pellizzer F, Crespi L, Arrigo G D, Lacaita A L and Servalli G 2014 *IEEE International Electron Devices Meeting*, San Francisco, CA, 2014, pp. 29.1.1–29.1.4

[76] Kim W, BrightSky M, Masuda T, Sosa N, Kim S, Bruce R, Carta F, Fraczak G, Cheng H Y, Ray A, Zhu Y, Lung H L, Suu K and Lam C 2016 *IEEE International Electron Devices Meeting (IEDM)*, San Francisco, CA, 2016, pp. 4.2.1–4.2.4

[77] Salinga M, Kersting B, Ronneberger I, Jonnalagadda V P, Vu X T, M Le Gallo, Giannopoulos I, Cojocaru-Miredin O, Mazzarello R and Sebastian A 2018 *Nat. Mater.* **17** 681

[78] Ding K, Wang J, Zhou Y, Tian H, Lu L, Mazzarello R, Jia C, Zhang W, Rao F and Ma E 2019 *Science* **366** 210

[79] Kuzum D, Jeyasingh R G, Lee B and Wong H S 2012 *Nano Lett.* **12** 2179

[80] Hosseini P, Sebastian A, Papandreou N, Wright C D and Bhaskaran H 2015 *IEEE Electron Device Lett.* **36** 975

[81] Cassinerio M, Ciocchini N and Ielmini D 2013 *Adv. Mater.* **25** 5975

[82] Loke D, Skelton J M, Wang W J, Lee T H, Zhao R, Chong T C and Elliott S R 2014 *Proc. Natl. Acad. Sci. USA* **111** 13272

[83] Gallo M L, Sebastian A, Cherubini G, Giefers H and Eleftheriou E 2017 *IEEE International Electron Devices Meeting (IEDM)*, San Francisco, CA, 2017, pp. 28.3.1–28.3.4

[84] Le Gallo M, Sebastian A, Cherubini G, Giefers H and Eleftheriou E 2018 *IEEE Transactions on Electron Devices* **65** 4304

[85] Sebastian A, Boybat I, Dazzi M, Giannopoulos I, Jonnalagadda V, Joshi V, Karunaratne G, Kersting B, Khaddam-Aljameh R, Nandakumar S R, Petropoulos A, Piveteau C, Antonakopoulos T, Rajendran B, Gallo M L and Eleftheriou E 2019 *Symposium on VLSI Technology*, Kyoto, Japan, 2019, pp. T168–T169

[86] Tuma T, Le Gallo M, Sebastian A and Eleftheriou E 2016 *IEEE Electron Device Lett.* **37** 1238

[87] Moraitis T, Sebastian A and Eleftheriou E 2018 *IEEE Nanotechnology Magazine* **12** 45

[88] Jian-Gang Z 2008 *Proc. IEEE* **96** 1786

[89] Slonczewski J C 1996 *J. Magn. Magn. Mater.* **159** L1

[90] Kawahara T, Ito K, Takemura R and Ohno H 2012 *Microelectronics Reliability* **52** 613

[91] Chappert C, Fert A and Van Dau F N 2007 *Nat. Mater.* **6** 813

[92] Locatelli N, Cros V and Grollier J 2014 *Nat. Mater.* **13** 11

[93] Hu G, Nowak J J, Gottwald M G, *et al.* 2019 *IEEE International Electron Devices Meeting (IEDM)*, San Francisco, CA, USA, 2019, pp. 2.6.1–2.6.4

[94] Carboni R, Ambrogio S, Chen W, Siddik M, Harms J, Lyle A, Kula W, Sandhu G and Ielmini D 2016 *IEEE International Electron Devices Meeting (IEDM)*, San Francisco, CA, 2016, pp. 21.6.1–21.6.4

[95] Yuasa S, Nagahama T, Fukushima A, Suzuki Y and Ando K 2004 *Nat. Mater.* **3** 868

[96] Lequeux S, Sampaio J, Cros V, Yakushiji K, Fukushima A, Matsumoto R, Kubota H, Yuasa S and Grollier J 2016 *Sci. Rep.* **6** 31510

[97] Raymenants E, Wan D, Couet S, Zografos O, Nguyen V D, Vaysset A, Souriau L, Thiam A, Manfrini M, Brus S, Heyns M, Mocuta D, Nikonov D E, Manipatruni S, Young I A, Devolder T and Radu I P 2018 *IEEE International Electron Devices Meeting (IEDM)*, San Francisco, CA, 2018, pp. 36.4.1–36.4.4

[98] Fukushima A, Seki T, Yakushiji K, Kubota H, Imamura H, Yuasa S and Ando K 2014 *Appl. Phys. Express* **7** 083001

[99] Chowdhury S, Datta S and Camsari K Y 2019 *IEEE International Electron Devices Meeting (IEDM)*, San Francisco, CA, USA, 2019, pp. 37.5.1–37.5.4

[100] Bez R, Camerlenghi E, Modelli A and Visconti A 2003 *Proc. IEEE* **91** 489

[101] Bayat F M, Guo X, Om'mani H A, Do N, Likharev K K and Strukov D B 2015 *IEEE International Symposium on Circuits and Systems (ISCAS)*, Lisbon, 2015, pp. 1921–1924

[102] Wang P, Xu F, Wang B, Gao B, Wu H, Qian H and Yu S 2019 *IEEE Transactions on Very Large Scale Integration (VLSI) Systems* **27** 988

[103] Guo X, Bayat F M, Bavandpour M, Klachko M, Mahmoodi M R, Prezioso M, Likharev K K and Strukov D B 2017 *IEEE International Electron Devices Meeting (IEDM)*, San Francisco, CA, 2017, pp. 6.5.1–6.5.4

[104] Kim M, Kim J, Park G, Everson L, Kim H, Song S, Lee S and Kim C H 2018 *IEEE International Electron Devices Meeting (IEDM)*, San Francisco, CA, 2018, pp. 15.4.1–15.4.4

[105] Lue H, Hsu P, Wei M, Yeh T, Du P, Chen W, Wang K and Lu C 2019 *IEEE International Electron Devices Meeting (IEDM)*, San Francisco, CA, USA, 2019, pp. 38.1.1–38.1.4

[106] Böcke T S, Müler J, Brähaus D, Schröer U and Bötger U 2011 *Appl. Phys. Lett.* **99** 102903

[107] Trentzsch M, Flachowsky S, Richter R, Paul J, Reimer B, Utess D, Jansen S, Mulaosmanovic H, Müler S, Slesazeck S, Ocker J, Noack M, Müler J, Polakowski P, Schreiter J, Beyer S, Mikolajick T and Rice B 2016 *IEEE International Electron Devices Meeting (IEDM)*, San Francisco, CA, 2016, pp. 11.5.1–11.5.4

[108] Yin X, Aziz A, Nahas J, Datta S, Gupta S, Niemier M and Hu X S 2016 *IEEE/ACM International Conference on Computer-Aided Design (ICCAD)*, Austin, TX, 2016, pp. 1–8

[109] Ni K, Grisafe B, Chakraborty W, Saha A K, Dutta S, Jerry M, Smith J A, Gupta S and Datta S 2018 *IEEE International Electron Devices Meeting (IEDM)*, San Francisco, CA, 2018, pp. 16.1.1–16.1.4

[110] Ni K, Smith J A, Grisafe B, Rakshit T, Obradovic B, Kittl J A, Rodder M and Datta S 2018 *IEEE International Electron Devices Meeting (IEDM)*, San Francisco, CA, 2018, pp. 13.2.1–13.2.4

[111] Luo J, Yu L, Liu T, Yang M, Fu Z, Liang Z, Chen L, Chen C, Liu S, Wu S, Huang Q and Huang R 2019 *IEEE International Electron Devices Meeting (IEDM)*, San Francisco, CA, USA, 2019, pp. 6.4.1–6.4.4

[112] Sun X, Wang P, Ni K, Datta S and Yu S 2018 *IEEE International Electron Devices Meeting (IEDM)*, San Francisco, CA, 2018, pp. 3.1.1–3.1.4

[113] Joshi V K and Lobo H C 2016 *Advances in Intelligent Systems and Computing*, pp. 25–40

[114] Jeloka S, Akesh N B, Sylvester D and Blaauw D 2016 *IEEE Journal of Solid-State Circuits* **51** 1009

[115] Li S, Niu D, Malladi K T, Zheng H, Brennan B and Xie Y 2017 *50th Annual IEEE/ACM International Symposium on Microarchitecture (MICRO)*, Boston, MA, USA, 2017, pp. 288–301

[116] Seshadri V, Lee D, Mullins T, Hassan H, Boroumand A, Kim J, Kozuch M A, Mutlu O, Gibbons P B and Mowry T C 2017 *50th Annual IEEE/ACM International Symposium on Microarchitecture (MICRO)*, Boston, MA, USA, 2017, pp. 273–287

[117] Dong Q, Sinangil M E, Erbagci B, Sun D, Khwa W, Liao H, Wang Y and Chang J 2020 *IEEE International Solid- State Circuits Conference - (ISSCC)*, San Francisco, CA, USA, 2020, pp. 242–244

[118] Hills G, Lau C, Wright A, Fuller S, Bishop M D, Sriman T i, Kanhaiya P, R Ho, A Amer, Y Stein, D Murphy, Arvind, Chandrakasan A and Shulaker M M 2019 *Nature* **572** 595

[119] Mennel L, Symonowicz J, Wachter S, Polyushkin D K, Molina-Mendoza A J and Mueller T 2020 *Nature* **579** 62

[120] Zhu X, Li D, Liang X and Lu W D 2019 *Nat. Mater.* **18** 141

[121] Zhu J, Yang Y, Jia R, Liang Z, Zhu W, Rehman Z U, Bao L, Zhang X, Cai Y, Song L and Huang R 2018 *Adv. Mater.* **30** 1800195

[122] Sangwan V K, Lee H S, Bergeron H, Balla I, Beck M E, Chen K S and Hersam M C 2018 *Nature* **554** 500

[123] Wang S, Chen C, Yu Z, He Y, Chen X, Wan Q, Shi Y, Zhang D W, Zhou H, Wang X and Zhou P 2019 *Adv. Mater.* **31** 1806227

[124] Xia Q, Robinett W, Cumbie M W, Banerjee N, Cardinali T J, Yang J J, Wu W, Li X, Tong W M, Strukov D B, Snider G S, Medeiros-Ribeiro G and Williams R S 2009 *Nano Lett.* **9** 3640

[125] Jo S H, Chang T, Ebong I, Bhadviya B B, Mazumder P and Lu W 2010 *Nano Lett.* **10** 1297

[126] Merolla P A, Arthur J V, Alvarez-Icaza R, *et al.* 2014 *Science* **345** 668

[127] Davies M, SrinivasNa, Lin T, *et al.* 2018 *IEEE Micro* **38** 82

[128] Sangwan V K and Hersam M C 2020 *Nat. Nanotechnol.*

[129] Sanchez Esqueda I, Yan X, Rutherglen C, Kane A, Cain T, Marsh P, Liu Q, Galatsis K, Wang H and Zhou C 2018 *ACS Nano* **12** 7352

[130] Pang Y, Gao B, Wu D, Yi S, Liu Q, W Chen, Chang T, Lin W, Sun X, Yu S, Qian H, Chang M and Wu H 2019 *IEEE International Solid-State Circuits Conference - (ISSCC)*, San Francisco, CA, USA, 2019, pp. 402–404

[131] Holcomb D E, Burleson W P and Fu K 2009 *IEEE Transactions on Computers* **58** 1198

[132] Alaghi A and Hayes J P 2013 *ACM Transactions on Embedded Computing Systems* **2** 1

[133] Gaba S, Sheridan P, Zhou J, Choi S and Lu W 2013 *Nanoscale* **5** 5872

[134] Huang C Y, Shen W C, Tseng Y H, King Y C and Lin C J 2012 *IEEE Electron Device Letters* **33** 1108

[135] Balatti S, Ambrogio S, Wang Z and Ielmini D 2015 *IEEE Journal on Emerging and Selected Topics in Circuits and Systems* **5** 214

[136] Jiang H, Belkin D, Savel'ev S E, Lin S, Wang Z, Li Y, S Joshi, Midya R, Li C, Rao M, Barnell M, Wu Q, Yang J J and Xia Q 2017 *Nat. Commun.* **8** 882

[137] Knag P, Gaba S, Lu W Zhang and Z 2019 *RRAM Solutions for Stochastic Computing. Stochastic Computing: Techniques and Applications*, pp. 153–164

[138] Borders W A, Pervaiz A Z, Fukami S, Camsari K Y, Ohno H and Datta S 2019 *Nature* **573** 390

[139] Brown B D and Card H C 2001 *IEEE Transactions on Computers* **50** 891

[140] Qian W, X Li, Riedel M D, Bazargan K and Lilja D J 2011 *IEEE Transactions on Computers* **60** 93

[141] Gaudet V C and Rapley A C 2003 *Electron. Lett.* **39** 299

[142] Amlani I, Orlov A O, Toth G, Bernstein G H, Lent C S and Snider G L 1999 *Science* **284** 289

[143] Khajetoorians A A, Wiebe J, Chilian B and Wiesendanger R 2011 *Science* **332** 1062

[144] Huang P, Kang J, Zhao Y, Chen S, Han R, Zhou Z, Chen Z, Ma W, Li M, Liu L and Liu X 2016 *Adv. Mater.* **28** 9758

[145] Kvatinsky S, Belousov D, Liman S, Satat G, Wald N, Friedman E G, Kolodny A and Weiser U C 2014 *IEEE Transactions on Circuits and Systems II: Express Briefs* **61** 895

[146] Talati N, Gupta S, Mane P and Kvatinsky S 2016 *IEEE Transactions on Nanotechnology* **15** 635

[147] Xiang Y C, Huang P, Yang H Z, Wang K L, Han R Z, Shen W S, Feng Y L, Liu C, Liu X Y and Kang J F 2019 *IEEE International Electron Devices Meeting (IEDM)*, San Francisco, CA, USA, 2019, pp. 38.2.1–38.2.4

[148] Sebastian A, Tuma T, Papandreou N, M Le Gallo, Kull L, Parnell T and Eleftheriou E 2017 *Nat. Commun.* **8** 1115

[149] Le Gallo M, Sebastian A, Mathis R, M Manica, Giefers H, Tuma T, Bekas C, Curioni A and Eleftheriou E 2018 *Nat. Electron.* **1** 246

[150] Bagdzevicius S, Maas K, Boudard M and Burriel M 2017 *Journal of Electroceramics* **39** 157

[151] Wang Z, Kang J, Fang Y, Yu Z, Yang X, Cai Y, Wang Y and Huang R 2016 *16th Non-Volatile Memory Technology Symposium (NVMTS)*, Pittsburgh, PA, 2016, pp. 1–3

[152] Choi S, Tan S H, Li Z, Kim Y, Choi C, Chen P Y, Yeon H, Yu Sand Kim J 2018 *Nat. Mater.* **17** 335

[153] Yang J J, Zhang M X, Strachan J P, Miao F, Pickett M D, Kelley R D, Medeiros-Ribeiro G and Williams R S 2010 *Appl. Phys. Lett.* **97** 232102

[154] Balatti S, Ambrogio S, Wang Z, Sills S, Calderoni A, Ramaswamy N and Ielmini D 2015 *IEEE Transactions on Electron Devices* **62** 3365

[155] Chen C Y, Goux L, Fantini A, Redolfi A, Clima S, Degraeve R, Chen Y Y, Groeseneken G and Jurczak M 2014 *IEEE International Electron Devices Meeting*, San Francisco, CA, 2014, pp. 14.2.1–14.2.4

[156] Cai Y, Lin Y, Xia L, Chen X, Han S, Wang Y and Yang H 2018 *55th ACM/ESDA/IEEE Design Automation Conference (DAC)*, San Francisco, CA, 2018, pp. 1–6

[157] Ishimaru K 2019 *IEEE International Electron Devices Meeting (IEDM)*, San Francisco, CA, USA, 2019, pp. 1.3.1–1.3.6

[158] Woo J, K Moon, Song J, Lee S, Kwak M, Park J and Hwang H 2016 *IEEE Electron Device Lett.* **37** 994

[159] Ambrogio S, Narayanan P, Tsai H, Shelby R M, Boybat I, di Nolfo C, Sidler S, M Giordano, Bodini M, Farinha N C P, Killeen B, Cheng C, Jaoudi Y and Burr G W 2018 *Nature* **558** 60
[160] Yoon K J, Bae W, Jeong D K and Hwang C S 2016 *Advanced Electronic Materials* **2** 1600326
[161] Burr G W, Shelby R M, Sidler S, C di Nolfo, Jang J, Boybat I, Shenoy R S, Narayanan P, Virwani K, Giacometti,E U Kurdi B N and Hwang H 2015 *IEEE Transactions on Electron Devices* **62** 3498
[162] Chen P, Lin B, Wang I, Hou T, Ye J, Vrudhula S, Seo J, Cao Y and Yu S 2015 *IEEE/ACM International Conference on Computer-Aided Design (ICCAD)*, Austin, TX, 2015, pp. 194–199
[163] Chen P, Peng X and Yu S 2017 *IEEE International Electron Devices Meeting (IEDM)*, San Francisco, CA, 2017, pp. 6.1.1–6.1.4
[164] Zidan M A, Fahmy H A H, Hussain M M and Salama K N 2013 *Microelectronics Journal* **44** 176
[165] Midya R, Wang Z, Zhang J, Savel'ev S E, Li C, Rao M, Jang M H, Joshi S, Jiang H, Lin P, Norris K, Ge N, Wu Q, Barnell M, Li Z, Xin H L, Williams R S, Xia Q and Yang J J 2017 *Adv. Mater.* **29** 1604457
[166] Woo J, Peng X and Yu S 2018 *IEEE International Symposium on Circuits and Systems (ISCAS)*, Florence, 2018, pp. 1–4
[167] Li C, Belkin D, Li Y, Yan P, Hu M, Ge N, Jiang H, Montgomery E, Lin P, Wang Z, Song W, Strachan J P, Barnell M, Wu Q, Williams R S, Yang J J and Xia Q 2018 *Nat. Commun.* **9** 2385
[168] Reuben J, Ben-Hur R, Wald N, Talati N, Ali A H, Gaillardon P and Kvatinsky S 2017 *27th International Symposium on Power and Timing Modeling*, Optimization and Simulation (PATMOS), Thessaloniki, 2017, pp. 1–8
[169] Chen P, Kadetotad D, Xu Z, Mohanty A, Lin B, Ye J, Vrudhula S, Seo J, Cao Y and Yu S 2015 *Design, Automation & Test in Europe Conference & Exhibition (DATE)*, 2015
[170] Xu C, Niu D, Muralimanohar N, Balasubramonian R, Zhang T, Yu S and Xie Y 2015 *IEEE 21st International Symposium on High Performance Computer Architecture (HPCA)*, Burlingame, CA, 2015, pp. 476–488
[171] Geim A K and Novoselov K S 2009 *The rise of graphene. Nanoscience and Technology* pp. 11–19
[172] Lin P, Li C, Wang Z, Li Y, Jiang H, Song W, Rao M, Zhuo Y, Upadhyay N K, Barnell M, Wu Q, Yang J J and Xia Q 2020 *Nat. Electron.* **3** 252
[173] Lin J, Xia L, Zhu Z, Sun H, Cai Y, Gao H, Cheng M, Chen X, Wang Y and Yang H 2018 *2018 Design, Automation & Test in Europe Conference & Exhibition (DATE)*, Dresden, 2018, pp. 407–412
[174] Zhu Z, Ma M, Liu J, Xu L, Chen X, Yang Y, Wang Y and Yang H 2019 *IEEE/ACM International Conference on Computer-Aided Design (ICCAD)*, Westminster, CO, USA, 2019, pp. 1–8
[175] Biswas A and Chandrakasan A P 2019 *IEEE Journal of Solid-State Circuits* **54** 217
[176] Xue C, Chen W, Liu J, *et al.* 2019 *IEEE International Solid- State Circuits Conference - (ISSCC)*, San Francisco, CA, USA, 2019, pp. 388–390
[177] Sun H, Zhu Z, Cai Y, Chen X, Wang Y and Yang H 2020 *25th Asia and South Pacific Design Automation Conference (ASP-DAC)*, Beijing, China, 2020, pp. 325–330
[178] Wan W, Kubendran R, Eryilmaz S B, Zhang W, Liao Y, Wu D, Deiss S, Gao B, Raina P, Joshi S, Wu H, Cauwenberghs G and Wong H P 2020 *IEEE International Solid- State Circuits Conference - (ISSCC)*, San Francisco, CA, USA, 2020, pp. 498–500
[179] Kim M, Liu M, Everson L, Park G, Jeon Y, Kim S, Lee S, Song S and Kim C H 2019 *IEEE International Electron Devices Meeting (IEDM)*, San Francisco, CA, USA, 2019, pp. 38.3.1–38.3.4
[180] Valavi H, Ramadge P J, Nestler E and Verma N 2019 *IEEE Journal of Solid-State Circuits* **54** 1789
[181] Lin B, Gao B, Pang Y, Yao P, Wu D, He H, Tang J, Qian H and Wu H 2019 *IEEE International Electron Devices Meeting* , San Francisco, CA, USA, 2019, pp. 14.8.1–14.8.4
[182] Mahmoodi M R, Nili H, Fahimi Z, Larimian S, Kim H and Strukov D 2019 *IEEE International Electron Devices Meeting (IEDM)*, San Francisco, CA, USA, 2019, pp. 30.1.1–30.1.4